

# Generalization in Machine Learning through Information-Theoretic Lens

by

Ziqiao Wang

Thesis submitted to the University of Ottawa  
in partial fulfillment of the requirements for the degree of  
Doctorate in Philosophy  
in  
Electrical and Computer Engineering

School of Electrical Engineering and Computer Science  
Faculty of Engineering  
University of Ottawa

© Ziqiao Wang, Ottawa, Canada, 2024

### **Examining Committee Membership**

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor: Yongyi Mao  
Professor  
School of Electrical Engineering and Computer Science  
University of Ottawa

Internal Member: Maia Fraser  
Associate Professor  
Department of Mathematics and Statistics  
University of Ottawa

Internal Member: Tommaso R. Cesari  
Assistant Professor  
School of Electrical Engineering and Computer Science  
University of Ottawa

Carleton Member: James R. Green  
Professor  
Department of Systems and Computer Engineering  
Carleton University

External Member: Ashish Khisti  
Professor  
Department of Electrical and Computer Engineering  
University of Toronto



### **Author's Declaration**

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Abstract

In this thesis, we utilize an information-theoretic framework to investigate generalization theory in machine learning, a critical area of research today. Specifically, we develop novel information-theoretic generalization bounds for machine learning algorithms. First, we apply information-theoretic analysis for models trained using stochastic gradient descent (SGD). We do so by invoking an auxiliary weight process and by approximating SGD using stochastic differential equations (SDE). Our analysis reveals intriguing phenomena such as epoch-wise double descent of gradient dispersion when trained with noisy labels. We also use our bounds to design new regularization techniques, including dynamic gradient clipping and Gaussian model perturbation, that can improve generalization performance. Furthermore, our framework is not limited to SGD-based algorithms; we also derive new information-theoretic bounds for any black-box learning algorithm, which are tighter than previous results based on the same settings. In addition, we apply our analysis to unsupervised domain adaptation (UDA), obtaining generalization bounds for two notions of the generalization error. Our algorithm-dependent bounds enable us to design new regularization techniques that can boost the performance of domain adaptation algorithms. Finally, we combine the stability-based generalization analysis with our information-theoretic analysis to derive novel generalization bounds, which explain the generalization in cases where previous information-theoretic bounds have fallen short.

*To Mom, Dad, and Xinyue*

## Acknowledgements

The past several years of my PhD career have been a wonderful journey. This journey has been supported tremendously from people I have encountered. Before I delve into the technical content of my thesis, I would like to thank all of them.

First and foremost, I would like to thank my amazing advisor, Yongyi Mao, for being the best type of PhD adviser any student could wish for. He is a brilliant researcher and always is full of enthusiasm for research. His way of thinking, in-depth knowledge, passion, and optimism guided me throughout my PhD journey. He has cultivated in me a taste for fundamental research problems, a gift that I believe I could never have acquired without his mentorship. I vividly recall one of our early conversations, during my time as a master's student, when we discussed artificial intelligence and machine learning. He remarked, "Communication is human encoding and human decoding; AI is God encoding and human decoding. In both cases, information theory is an indispensable tool." Additionally, at the beginning of my PhD journey, he shared his vision by stating, "I believe that learning theory can be rewritten through the lens of information theory." While, to be honest, I struggled to grasp the significance of these big visions initially, his unwavering conviction and willingness to challenge conventional wisdom left a lasting impression on me, and today, the title of this thesis stands as a confirmation to the validity of his vision, at least in part. His support extends to every aspect of my PhD career. For instance, he consistently encourages and backs my participation in academic conferences and workshops, even when I have no work to present. One memorable instance was when he suggested I attend a deep learning theory workshop at IAS in Princeton in 2019. Despite initial concerns about my mathematical maturity, he believed it would be a valuable experience for me. Indeed, that workshop turned out to be a crucial experience in my early PhD career. Although I could not understand the majority of talks at the time, I was deeply attracted to the theoretical works presented in that workshop. It was during this event that I realized the appeal of fundamental research. Upon my return to Ottawa, I resolved to focus my PhD research on the topic of generalization. These moments from my PhD journey are truly unforgettable, and undoubtedly, being guided by Prof. Yongyi Mao has been a tremendous honor throughout this remarkable experience.

I am grateful to have Maia Fraser, Tom Cesari, James Green, and Ashish Khisti as my thesis committee members, and I extend particular thanks to each of them for nominating my PhD thesis for the university-level thesis medal and prize. I express my gratitude to Prof. Fraser for her unwavering support throughout my PhD journey and for serving as a reference for my job applications. I am thankful to Prof. Green for his careful reading of

this thesis and for providing invaluable feedback. Additionally, I appreciate Prof. Cesari and Prof. Khisti for their encouraging words regarding my PhD thesis work.

I would like to thank the members of YM's group, including Masoumeh Soflaei, Guillaume P. Archambault, Dwarak Govind Parthiban, Jie Meng, Chune Li, Xing Xu, Runzhi Tian, Boming Shi, Chenjie Sun, Junshuang Wu, Siyuan Yang, Zixuan Liu, Zhiyi Dong, Alim Manjiyani, Hongyi Xiong, Sammy Hannat, and many others. It has been a pleasure to have the company of all of you during my time at uOttawa, and I will always cherish the memories of the gatherings and parties we shared together.

I would like to express my gratitude to several senior researchers who have provided invaluable advice and support, both in terms of my research career and personal life. I am particularly thankful to Maia Fraser, Harry H.Y. Guo, Ali Al-Bashabsheh, Richong Zhang, Yong Liu, Yuheng Bu, and Gergely Neu. In particular, I thank Harry for hosting me when I was an intern at National Research Council Canada and for invaluable guidance on navigating the challenges of being a PhD student. I also extend my thanks to Prof. Zhang and Prof. Liu for providing me numerous opportunities to present my research among peers in China and for giving sincere guidance regarding my future research career. I appreciate Prof. Neu for identifying some issues in my first PhD research work, which greatly contributed to my professional growth during the early stages of my career. Additionally, I am thankful to Prof. Bu for his encouraging words regarding my research and for agreeing to serve as a reference for my job applications.

I would like to thank many young researchers whom I've had the pleasure of meeting, including Xingxiang Li, Yuan Wu, Ruikun Zhou, Chengyu Dong, Weisen Jiang, Yuchen Li, Mahdi Haghifam, Fredrik Hellström, Borja Rodríguez Gálvez, Haiyun He, Hao Wang, Ruida Zhou, Zhanke Zhou, Zhengyu Zhang, Xiaolin Hu, Zhijie Nie, Fanshuang Kong and so on. Whether working on similar research topics or not, every discussion about research with you guys has been enriching and enjoyable. Furthermore, I want to express my appreciation to my collaborators, Zixuan Liu, Fanshuang Kong, Zhijie Nie, and Hailang Huang, for collaborating on various stimulating projects together.

I would like to thank my friends in Ottawa who have supported me and helped me navigate life in Canada, especially during the challenges posed by the pandemic. Special thanks to Weiwen Raymon Lin, Runzhi Tian, Junhao Yan, Fan Yang, Lingfeng Zhang, Ao Peng, Chenjie Sun, Xingxiang Li, Zixiong Han, Jiadi Wu, Hao Li, Yongwei Yuan, Chen Duan, Bowen Gu, Zijian Long, Sijun Mao and so on. Having all of you by my side during my time in Ottawa means a great deal to me, and I will never forget every lunch or dinner party, every board role-playing game, every card game, every gym session and every basketball game we shared together. Furthermore, I wish to express my gratitude to

my high school deskmate, Judong Ma. Thank you for staying in touch over the years and for sharing both your joys and sorrows with me. Despite the physical distance between us, your friendship has remained deeply meaningful to me.

Lastly, I extend my deepest appreciation to my parents, Ping Bao and Lilai Wang, for their unwavering support. Without their encouragement, this thesis would not have been possible. Additionally, I want to express my heartfelt appreciation to Xinyue, my best friend and the love of my life since my undergraduate years. You have been the driving force behind my decision to study abroad, and your presence in my life has made me a better person than I ever thought possible. Maintaining a long-distance relationship has undoubtedly been challenging, and mere words cannot adequately convey my gratitude for your understanding and support. I still cherish the words you shared with me by the Huangpu River, urging me to pursue my passions and stay true to myself. Your constant encouragement has been a beacon of light during the ups and downs of my PhD journey, and I attribute much of my achievements to your steadfast presence in my life. I am filled with excitement as I look forward to the bright future that awaits us.

# Table of Contents

|                                                           |          |
|-----------------------------------------------------------|----------|
| Examining Committee                                       | ii       |
| Author's Declaration                                      | iii      |
| Abstract                                                  | iv       |
| Dedication                                                | v        |
| Acknowledgements                                          | vi       |
| List of Figures                                           | xiv      |
| List of Tables                                            | xvii     |
| <b>1 Introduction</b>                                     | <b>1</b> |
| 1.1 Background . . . . .                                  | 1        |
| 1.2 Dissertation Overview . . . . .                       | 4        |
| <b>2 Notations and Preliminaries</b>                      | <b>7</b> |
| 2.1 Probability and Information Theory Notation . . . . . | 7        |
| 2.2 Generalization Error . . . . .                        | 7        |
| 2.3 Other Prerequisite Definitions . . . . .              | 8        |
| 2.4 Technical Lemmas . . . . .                            | 9        |

|          |                                                                                                                                       |           |
|----------|---------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>3</b> | <b>On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications</b>                                | <b>12</b> |
| 3.1      | Overview and Main Contributions . . . . .                                                                                             | 12        |
| 3.2      | Other Related Literature . . . . .                                                                                                    | 14        |
| 3.3      | Additional Preliminaries . . . . .                                                                                                    | 15        |
| 3.4      | New Generalization Bounds for SGD . . . . .                                                                                           | 17        |
| 3.5      | Application in Linear Networks and Two-Layer ReLU Networks . . . . .                                                                  | 21        |
| 3.6      | Experimental Study . . . . .                                                                                                          | 22        |
| 3.7      | A Practical Implication: Gaussian Model Perturbation . . . . .                                                                        | 25        |
| 3.8      | Conclusion . . . . .                                                                                                                  | 26        |
| <b>4</b> | <b>Two Facets of SDE Under an Information-Theoretic Lens: Generalization of SGD via Training Trajectories and via Terminal States</b> | <b>28</b> |
| 4.1      | Overview and Main Contributions . . . . .                                                                                             | 28        |
| 4.2      | Additional Preliminaries . . . . .                                                                                                    | 31        |
| 4.3      | Generalization Bounds Via Full Trajectories . . . . .                                                                                 | 33        |
| 4.4      | Generalization Bounds Via Terminal State . . . . .                                                                                    | 35        |
| 4.5      | Empirical Study . . . . .                                                                                                             | 41        |
| 4.6      | Conclusion . . . . .                                                                                                                  | 43        |
| <b>5</b> | <b>Tighter Information-Theoretic Generalization Bounds from Supersamples</b>                                                          | <b>45</b> |
| 5.1      | Overview and Main Contributions . . . . .                                                                                             | 45        |
| 5.2      | Supersample Setting . . . . .                                                                                                         | 46        |
| 5.3      | Generalization Bounds via Loss Difference . . . . .                                                                                   | 46        |
| 5.4      | Generalization Bounds via Correlating with Rademacher Sequence . . . . .                                                              | 50        |
| 5.5      | Numerical Results . . . . .                                                                                                           | 57        |
| 5.6      | Limitations and Other Related Literature . . . . .                                                                                    | 60        |
| 5.7      | Conclusion . . . . .                                                                                                                  | 61        |



|          |                                                                                                     |            |
|----------|-----------------------------------------------------------------------------------------------------|------------|
| <b>6</b> | <b>Information-Theoretic Analysis of Unsupervised Domain Adaptation</b>                             | <b>62</b>  |
| 6.1      | Overview and Main Contributions . . . . .                                                           | 62         |
| 6.2      | Related Work . . . . .                                                                              | 63         |
| 6.3      | Refined Notation Systems . . . . .                                                                  | 64         |
| 6.4      | Upper Bounds for PP Generalization Error . . . . .                                                  | 65         |
| 6.5      | Upper Bounds for EP Generalization Error and Applications . . . . .                                 | 69         |
| 6.6      | Experimental Results . . . . .                                                                      | 74         |
| 6.7      | Conclusion . . . . .                                                                                | 75         |
| <b>7</b> | <b>Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds</b> | <b>76</b>  |
| 7.1      | Overview and Main Contributions . . . . .                                                           | 76         |
| 7.2      | Other Related Literature . . . . .                                                                  | 79         |
| 7.3      | Additional Preliminaries . . . . .                                                                  | 79         |
| 7.4      | IOMI Bounds for Stable Algorithms . . . . .                                                         | 82         |
| 7.5      | CMI Bounds for Stable Algorithms . . . . .                                                          | 83         |
| 7.6      | Convex–Lipschitz–Bounded (CLB) Problems . . . . .                                                   | 86         |
| 7.7      | Extensions . . . . .                                                                                | 88         |
| 7.8      | Related Works and Additional Discussions . . . . .                                                  | 90         |
| 7.9      | Conclusion . . . . .                                                                                | 92         |
| <b>8</b> | <b>Conclusion and Future Work</b>                                                                   | <b>94</b>  |
|          | <b>References</b>                                                                                   | <b>97</b>  |
|          | <b>APPENDICES</b>                                                                                   | <b>121</b> |

|          |                                                                                               |            |
|----------|-----------------------------------------------------------------------------------------------|------------|
| <b>A</b> | <b>Omitted Proofs and Results in Chapter 3</b>                                                | <b>122</b> |
| A.1      | HWI Inequality: Proof of Lemma 3.4.1 . . . . .                                                | 122        |
| A.2      | A Key Lemma: Lemma 3.4.2 . . . . .                                                            | 123        |
| A.3      | Mutual Information Bounds for SGD: Theorem 3.4.2 . . . . .                                    | 124        |
| A.4      | Application in Neural Networks: Proof of Theorems 3.5.1 and 3.5.2 . . . . .                   | 130        |
| A.5      | Experiment Details . . . . .                                                                  | 132        |
| <b>B</b> | <b>Omitted Proofs and Results in Chapter 4</b>                                                | <b>141</b> |
| B.1      | Additional Background . . . . .                                                               | 141        |
| B.2      | Omitted Proofs and Additional Results in Section 4.3 . . . . .                                | 145        |
| B.3      | Omitted Proofs and Additional Results in Section 4.4 . . . . .                                | 150        |
| B.4      | Experiment Details and Additional Results . . . . .                                           | 155        |
| B.5      | Additional Result: Inverse Population FIM as both Posterior and Prior<br>Covariance . . . . . | 155        |
| <b>C</b> | <b>Omitted Proofs and Results in Chapter 5</b>                                                | <b>159</b> |
| C.1      | Some Additional Definitions on Chaining Methods . . . . .                                     | 159        |
| C.2      | Omitted Proofs and Additional Results in Section 5.3 . . . . .                                | 160        |
| C.3      | Omitted Proofs and Additional Results in Section 5.4 . . . . .                                | 166        |
| C.4      | Some Background on Channel Capacity of Binary Channel . . . . .                               | 176        |
| C.5      | Experimental Details and Additional Results . . . . .                                         | 177        |
| <b>D</b> | <b>Omitted Proofs and Results in Chapter 6</b>                                                | <b>185</b> |
| D.1      | Some Additional Discussions and Clarifications . . . . .                                      | 185        |
| D.2      | Omitted Proofs and Additional Results in Section 6.4 . . . . .                                | 185        |
| D.3      | Omitted Proofs and Additional Discussions in Section 6.5 . . . . .                            | 192        |
| D.4      | Experiment Details . . . . .                                                                  | 201        |

|          |                                                                    |            |
|----------|--------------------------------------------------------------------|------------|
| <b>E</b> | <b>Omitted Proofs and Results in Chapter 7</b>                     | <b>207</b> |
| E.1      | Useful Facts . . . . .                                             | 207        |
| E.2      | Further Elaborations on SCH Stability . . . . .                    | 209        |
| E.3      | Omitted Proofs and Additional Discussions in Section 7.4 . . . . . | 210        |
| E.4      | Omitted Proofs in Section 7.5 . . . . .                            | 213        |
| E.5      | Omitted Proof in Section 7.7 . . . . .                             | 222        |
| E.6      | CLB Examples . . . . .                                             | 223        |
| E.7      | Additional Applications . . . . .                                  | 225        |

# List of Figures

|     |                                                                                                                                                                                                                                        |    |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Comparison of the trajectory term between our bound (Theorem 3.4.2) and the bound in [153]. (a)(b) MLP trained on MNIST. (c)(d) CNN trained on CIFAR-10. . . . .                                                                       | 14 |
| 3.2 | Estimated bound and empirical generalization gap (“gap”) as functions of network width ((a) and (b)) and label noise level ((c) and (d)). Note that Y-axis is in log scale. . . . .                                                    | 23 |
| 3.3 | Dynamics of gradient dispersion, in relation to training/testing accuracies. . . . .                                                                                                                                                   | 24 |
| 3.4 | Dynamic Gradient Clipping. . . . .                                                                                                                                                                                                     | 24 |
| 4.1 | Performance of VGG-11 and ResNet-18 trained with SGD and SDE. Standard data augmentation techniques are only used in (d). . . . .                                                                                                      | 29 |
| 4.2 | Gradient-related quantities of SGD or its discrete SDE approximation. In (d), since per-sample gradient is ill-defined when BatchNormalization is used, we do not track $tr \{ \log (\Sigma_t^{-1} \Sigma_t^\mu) \}$ . . . . .         | 36 |
| 4.3 | Hessian-related quantities of SGD or its discrete SDE approximation. . . . .                                                                                                                                                           | 37 |
| 4.4 | (a-b) The dynamics of $\eta/2 - \lambda_1$ . Note that learning rate decays by 0.1 at the 40,000 <sup>th</sup> and the 60,000 <sup>th</sup> iteration. (c-d) The distance of current model parameters from its initialization. . . . . | 40 |
| 4.5 | Estimated trajectories-based bound and terminal-state based bound, with $R$ excluded. Zoomed-in figures of generalization error are given in Figure B.2 in Appendix. . . . .                                                           | 43 |
| 5.1 | Left: channel from $U_i$ to $\Delta L_i$ . Right: channel from $U_i$ to $L_i^+$ . Zero-one loss assumed. . . . .                                                                                                                       | 48 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |     |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.2 | Comparison of bounds on the synthetic eCMI/dataset. (a) Binary classification with a separable $\mu$ (i.e. the interpolating setting). Notice that the variance bound nearly coincides with $\mathcal{E}_\mu(\mathcal{A})$ . (b) Binary classification with a non-separable $\mu$ . (c) Ten-class classification with a separable $\mu$ . (d) Ten-class classification with a non-separable $\mu$ . The binary KL bound is removed in (d) since it is always $\geq 1$ . . . . .                                               | 56  |
| 5.3 | Comparison of bounds on two real datasets, MNIST (“4 vs 9”) and CIFAR10. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                              | 58  |
| A.1 | Comparison Between Theorem 3.4.2 and Lemma 3.3.1 in stochastic setting. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                               | 133 |
| A.2 | The impact of learning rate and batch size on the trajectory term and the flatness term in Eq. (3.4). . . . .                                                                                                                                                                                                                                                                                                                                                                                                                 | 135 |
| A.3 | Generalization gap of a linear network v.s. Theorem 3.5.1. Note y-axis is log-scale. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                  | 136 |
| A.4 | Generalization gap of a two-layer ReLU network v.s. Theorem 3.5.2. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 136 |
| A.5 | Dynamic Gradient Clipping (AlexNet). . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 137 |
| B.1 | (a-b) The dynamics of $\eta/2 - \lambda_1$ . Note that learning rate decays by 0.1 at the 40,000 <sup>th</sup> and the 60,000 <sup>th</sup> iteration. (c-d) The distance of current model parameters from its initialization. . . . .                                                                                                                                                                                                                                                                                        | 156 |
| B.2 | Zoomed-in of generalization error. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 157 |
| B.3 | Estimated trajectories-based bound and terminal-state based bound, with $R$ excluded. Models trained on CIFAR 10. . . . .                                                                                                                                                                                                                                                                                                                                                                                                     | 157 |
| C.1 | Comparison of $I(L_i^+; U_i)$ and $I(L_i^-; U_i)$ . . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 178 |
| C.2 | (a)(b) Comparison of the interpolating bounds on the synthetic dataset. Here <i>Channel Capacity</i> , <i>SIB</i> and <i>PIB</i> refer to the channel capacity result in Theorem 5.3.3, the single-loss based interpolating bound of Eq. (5.3) in Theorem 5.4.3 and a improved loss pair based interpolating bound of [92, Eq. (12)] (replacing $I(L_i; U_i   \tilde{Z})$ by $I(L_i; U_i)$ ), respectively. (c)(d) Further compare the value of $2I(L_i^+; U_i)$ and $I(L_i; U_i)$ , where training loss is non-zero. . . . . | 179 |
| C.3 | Verification of Theorem 5.4.2. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 180 |
| C.4 | Comparison of the square-root bounds on the synthetic dataset. Here <i>f-CMI</i> , <i>e-CMI</i> and <i>se-CMI</i> refer to the disintegrated <i>f</i> -CMI bound [82], the unconditional e-CMI bound in Corollary and the single-loss based square-root bound in Theorem 5.4.1, respectively. . . . .                                                                                                                                                                                                                         | 181 |

|     |                                                                                                                                                                                                                                                                                                       |     |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| C.5 | Comparison of three fast-rate bounds on the synthetic dataset. Here <i>Fast-Rate</i> refers to the fast-rate bound of Eq. (5.1) in Theorem 5.4.3. . . . .                                                                                                                                             | 183 |
| C.6 | Comparison of Optimal Fast-Rate Bounds. . . . .                                                                                                                                                                                                                                                       | 184 |
| D.1 | The relationship between random variables in UDA, where $F = L_{\mu'}(W) - L_S(W)$ . . . . .                                                                                                                                                                                                          | 186 |
| D.2 | Visualization results of representations obtained by using t-SNE. The source domain (blue points) is SVHN and the target domain (red points) is MNIST.                                                                                                                                                | 204 |
| D.3 | <b>KL</b> on <b>S</b> → <b>M</b> . The left figure is the comparison of the Jeffrey's divergence in the representation space and the testing error. The right figure is the evolution of testing accuracy with respect to the different fraction of unlabelled target data used for training. . . . . | 206 |

# List of Tables

|     |                                                                                                                                                                                                                                                 |     |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.1 | Top-1 classification accuracy acc.(%) of VGG16. We run experiments 10 times and report the mean and the standard deviation of the testing accuracy. Superscript denotes the number of sampled Gaussian noises during training. . . . .          | 26  |
| 6.1 | RotatedMNIST and Digits. Results of baselines are reported from [158]. . . . .                                                                                                                                                                  | 74  |
| 7.1 | Supersample (Left) and the induced hypotheses (Right). . . . .                                                                                                                                                                                  | 80  |
| A.1 | Top-1 classification accuracy acc.(%) of PreActResNet18. We run experiments 10 times and report the mean and the standard deviation of the testing accuracy. Superscript denotes the number of sampled Gaussian noises during training. . . . . | 139 |
| A.2 | Top-1 classification accuracy acc.(%) of VGG16. . . . .                                                                                                                                                                                         | 139 |
| B.1 | Comparison of the results in this work . . . . .                                                                                                                                                                                                | 142 |
| D.1 | RotatedMNIST and Digits Experiments of <b>ERM-CL</b> . Results of ERM are reported from [158]. . . . .                                                                                                                                          | 204 |
| D.2 | Ablation study on Effect of Gradient Penalty Hyperparameter. . . . .                                                                                                                                                                            | 204 |
| D.3 | VisDA17 experiments. Results of baselines are reported directly from [158].                                                                                                                                                                     | 205 |

# Chapter 1

## Introduction

Using information-theoretic generalization bounds to analyze the generalization properties of a learning algorithm has attracted increasing attention since the seminal works of [183, 184, 232]. One major advantage of such bounds is that the information-theoretic quantities, e.g., the mutual information (MI) between the training sample and the trained parameter weights, are both distribution-dependent and algorithm-dependent. This makes them an ideal tool to characterize the generalization properties of a learning algorithm, particularly when the traditional algorithm-independent learning-theoretic tools (e.g., VC-dimension [206] and Rademacher complexity [15]) appear inadequate. For example, [238] shows that the high-capacity deep neural networks still generalize well, contradicting the traditional wisdom in statistical learning theory that suggests complex models tend to overfit the training data and perform poorly on unseen data [206]. In contrast, the information-theoretic generalization bounds have experimentally demonstrated that they are capable of tracking the generalization behaviour of modern neural networks [151, 210, 82, 213, 214, 92].

### 1.1 Background

**Generalization in Modern Deep Learning** The observation that high capacity deep neural networks trained with mini-batched stochastic gradient descent, referred to SGD in this thesis, tend to generalize well [238] contradicts the classical wisdom in statistical learning theory (e.g., [206] ) and has stimulated intense research interest in understanding the generalization behaviour of modern neural networks.

In this direction, generalization bounds for over-parameterized neural networks are obtained [5, 14, 157, 155, 156, 9, 8, 130]. Additionally, a curious phenomenon known



as “double descent” has been observed and analyzed [19, 149, 234], where the testing error initially decreases, then increases, before experiencing a second decrease. Built on a connection between stability and generalization [24], a stability-based bound is first presented in [80], followed by a surge of research effort exploiting similar approaches [133, 36, 55, 118, 16].

Another line of research connects the generalization of neural networks with the flatness of loss minima [93] found by SGD or its variant [110, 50, 51, 154, 34, 100, 102, 244, 56]. This understanding has led to the discovery of new SGD-based training algorithms for improved generalization. For example, in a concurrent development by [244] and [56], a local “max-pooling” operation is applied to the loss landscape prior to the SGD updates. This approach, referred to as AMP [244] or SAM [56], is shown to make SGD favor flatter minima and achieve the state-of-the-art performance among various competitive regularization schemes.

**Unsupervised Domain Adaptation** Deep learning has achieved great success in many practical applications. One of the key assumptions in many deep learning scenarios, either in theory or in application, is that the training data and the testing data come from the same distribution. Notably, such an assumption does not always hold in practice, and the performance of trained models on the shifted distribution may behave significantly differently than on the original distribution. In this thesis, we will investigate the *unsupervised domain adaptation (UDA)* task, where the learner is confronted with a source domain and a target domain, and the algorithm is permitted to access labeled training samples from the source domain and unlabeled training samples from the target domain. The goal of this task is to find a predictor that performs well on the target domain.

A main obstacle in such a task is the discrepancy between the two domains. Some recent works [21, 20, 138, 243, 242, 190, 63, 3, 158] have proposed various measures to quantify such discrepancy, either for the UDA setting or for the more general domain generalization tasks, and many learning algorithms are proposed. For example, [158] uses a (reverse) KL divergence to measure the misalignment of the two domain distributions, and motivated by their generalization bound, they design an algorithm that penalizes the KL divergence between the marginal distributions of two domains in the representation space. Despite that this “KL guided domain adaptation” algorithm is demonstrated to outperform many existing marginal alignment algorithms [60, 200, 190, 120], it is not clear whether KL-based alignment of marginal distributions is adequate for UDA, and more fundamentally, what role the unlabelled target-domain sample should play in cross-domain generalization. Notably, most UDA algorithms are heuristically designed and intuitively justified. Moreover, most existing generalization bounds are algorithm-independent. Then there appears

significant room for both deeper theoretical understanding and more principled algorithm design.

**Information-Theoretic Generalization Framework** Information-theoretic upper bounds for generalization error have recently been developed since the seminal works of [183, 232]. On one hand, these bounds have attracted increasing interest due to their distribution dependence and algorithm dependence making them highly suitable for reasoning the generalization behavior of modern deep learning. In fact, subsequent studies have demonstrated that information-theoretic generalization bounds can effectively track the dynamics of generalization error in deep neural networks [151, 210, 82, 213, 214, 92, 217]. On the other hand, recent studies [199, 75, 82, 76, 92] have revealed the expressive nature of information-theoretic generalization bounds in the distribution-free setting.

The original information-theoretic generalization bound of [232] has been extended or improved in many different ways, such as the chaining method [11, 74, 248, 37], the subset or individual technique [151, 29, 77, 179, 247] and so on. Remarkably, [199] has developed generalization bounds based on a conditional mutual information (CMI) measure obtained for a “supersample” setting. Specifically, in addition to the original training sample, a “ghost sample” is introduced. The supersample is then constructed by combining these two sets, resulting in an  $n \times 2$  matrix of data instances. In each row, one instance is selected at random for training and the other masked out for testing. The authors then show that the CMI of the mask variables and the learned weights conditioned on the supersample can be used to upper-bound the generalization error. Additionally, in the case of the 0 – 1 loss and the realizable setting, where an interpolating algorithm attains zero empirical risk, CMI bounds give an exact characterization of the generalization error [76, 217], thereby providing the tightest possible generalization bound.

Although better behaving than the unconditional weight-based MI bounds (e.g., having boundedness guaranty), the CMI bounds can be difficult to measure for high-dimensional weights, which limits their application. To overcome such difficulty, functional CMI ( $f$ -CMI) bounds are proposed by [82], where the weight variable in CMI is replaced by the predictions for the supersample. In this case, each prediction pair is a two-dimensional discrete random variable, making the CMI easier to measure and also a tighter bound. More recently, [92] uses loss pairs to replace the predictions in  $f$ -CMI and obtain even tighter CMI bounds, known as evaluated CMI (e-CMI) bounds. In fact, the earliest version of e-CMI bound appeared in [199]. The notion was also exploited in later works [75, 76, 78]. Note that e-CMI still measures the dependence between a one-dimensional variable (mask) and a two-dimensional variable (loss pair). In this thesis, we show that it is possible

to further tighten the CMI bounds, using MI terms involving only two one-dimensional variables.

It is important to acknowledge that mutual information is just one option for quantifying the generalization of a learning algorithm. One compelling reason why mutual information-based metrics have garnered significant attention recently is their possession of several desirable properties, including the chain rule, the data processing inequality, and more. Throughout this thesis, we will explore various applications of these favorable properties.

**Information-theoretic generalization bounds for SGLD and SGD** Information-theoretic bounding techniques have proven to be powerful tools in analyzing SGD-like algorithms. For example, [168] was the first to utilize information-theoretic generalization bounds to analyze the generalization ability of stochastic gradient Langevin dynamics (SGLD) [62, 220]. This bound was subsequently improved upon by [151, 77, 179, 210]. To provide more accurate bounds, these works proposed data and algorithm-dependent quantities as the primary components of the bounds, such as gradient incoherence [151] and gradient variance [210]. Since the global Lipschitz parameter of loss used in [168] was too pessimistic, these works improved upon the bound by incorporating these data-dependent and algorithm-dependent quantities.

Furthermore, [153] was inspired by the work of [168] and presented an information-theoretic analysis of models trained with SGD. The analysis in [153] constructs an auxiliary weight process parallel to the SGD training process and upper-bounds the generalization error through this auxiliary process.

## 1.2 Dissertation Overview

The rest of this dissertation is structured as follows, where we also provide elaborated details to offer a comprehensive understanding of the current contributions (Chapter 3-7).

- In Chapter 2, we present the necessary notations, technical definitions, and prerequisite lemmas to facilitate understanding of the subsequent chapters.
- In Chapter 3, we follow up on a recent work of [153] and presents some new information-theoretic upper bounds for the generalization error of machine learning models, such as neural networks, trained with SGD. We apply these bounds to analyzing the generalization behaviour of linear and two-layer ReLU networks. Experimental study of

these bounds provide some insights on the SGD training of neural networks. They also point to a new and simple regularization scheme which we show performs comparably to the current state of the art.

This chapter is based on the work of [213]:

- Ziqiao Wang and Yongyi Mao, “On the Generalization of Models Trained with SGD: information-theoretic generalization bounds and Implications”, *International Conference on Learning Representations (ICLR) 2022*.
- In Chapter 4, we focus on the Stochastic differential equation (SDE) approximation of SGD, which have been shown recently to well characterize the dynamics of training machine learning models with SGD. This provides two opportunities for better understanding the generalization behaviour of SGD through its SDE approximation. Firstly, viewing SGD as full-batch gradient descent with Gaussian gradient noise allows us to obtain trajectories-based generalization bound using the information-theoretic generalization bound from [232]. Secondly, assuming mild conditions, we estimate the steady-state weight distribution of SDE and use information-theoretic generalization bounds from [232] and [151] to establish terminal-state-based generalization bounds. Our proposed bounds have some advantages, notably the trajectories-based bound outperforms results in [213], and the terminal-state-based bound exhibits a fast decay rate comparable to stability-based bounds.

This chapter is based on the work of [214]:

- Ziqiao Wang and Yongyi Mao, “Two Facets of SDE Under an Information-Theoretic Lens: Generalization of SGD via Training Trajectories and via Terminal States”, *Under Review (Short version presented at NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning (M3L))*.
- In Chapter 5, we present a variety of novel information-theoretic generalization bounds for learning algorithms, from the supsample setting of [199]—the setting of the “conditional mutual information” framework. Our development exploits projecting the loss pair (obtained from a training instance and a testing instance) down to a single number and correlating loss values with a Rademacher sequence (and its shifted variants). The presented bounds include square-root bounds, fast-rate bounds, including those based on variance and sharpness, and bounds for interpolating algorithms etc. We show theoretically or empirically that these bounds are tighter than all information-theoretic generalization bounds known to date on the same supsample setting.

This chapter is based on the work of [217]:

- Ziqiao Wang and Yongyi Mao, “Tighter Information-Theoretic Generalization Bounds from Supersamples”, *International Conference on Machine Learning (ICML) 2023*.
- In Chapter 6, we use information-theoretic tools to analyze the generalization error in unsupervised domain adaptation (UDA). We present novel upper bounds for two notions of generalization errors. The first notion measures the gap between the population risk in the target domain and that in the source domain, and the second measures the gap between the population risk in the target domain and the empirical risk in the source domain. While our bounds for the first kind of error are in line with the traditional analysis and give similar insights, our bounds on the second kind of error are algorithm-dependent, which also provide insights into algorithm designs. Specifically, we present two simple techniques for improving generalization in UDA and validate them experimentally.

This chapter is based on the work of [215]:

- Ziqiao Wang and Yongyi Mao, “Information-Theoretic Analysis of Unsupervised Domain Adaptation”, *International Conference on Learning Representations (ICLR) 2023*.
- In Chapter 7, we present new information-theoretic generalization guarantees through a novel construction of the “neighboring-hypothesis” matrix and a new family of stability notions termed sample-conditioned hypothesis (SCH) stability. Our approach offers sharper bounds that improve upon previous information-theoretic generalization bounds in various learning scenarios. Notably, these bounds address the limitations of existing information-theoretic generalization bounds in the context of stochastic convex optimization (SCO) problems, as explored in the recent work by [78].

This chapter is based on the work of [216]:

- Ziqiao Wang and Yongyi Mao, “Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds”, *Advances in Neural Information Processing Systems (NeurIPS) 2023*.

# Chapter 2

## Notations and Preliminaries

### 2.1 Probability and Information Theory Notation

Unless otherwise noted, a random variable will be denoted by a capitalized letter, and its realization by the corresponding lower-case letter. Let  $P_X$  denote the distribution of a random variable  $X$  and let  $P_{X|Y}$  be the conditional distribution of  $X$  conditioned on  $Y$ , which, upon conditioning on a specific realization  $y$ , is denoted by  $P_{X|Y=y}$  or simply  $P_{X|y}$ . Similarly,  $\mathbb{E}_X$  is the expectation taken over  $X \sim P_X$  and  $\mathbb{E}_{X|Y=y}$  (or  $\mathbb{E}_{X|y}$ ) is the expectation taken over  $X \sim P_{X|Y=y}$ . Let  $H(\cdot)$  be the entropy and let  $D_{\text{KL}}(P||Q)$  denote the KL divergence of  $P$  with respect to  $Q$ . Let  $I(X;Y)$  be the mutual information (MI) between  $X$  and  $Y$ , and  $I(X;Y|Z)$  the conditional mutual information between  $X$  and  $Y$  conditioned on  $Z$ . Also, we use  $\mathbb{W}(\cdot, \cdot)$  to denote the Wasserstein distance (formal definition is given in Appendix). Throughout the thesis, logarithm takes base  $e$ , making the unit of mutual information *nat*.

### 2.2 Generalization Error

Let  $\mathcal{Z}$  be the instance space of interest and  $\mu$  be an unknown distribution on  $\mathcal{Z}$ , specifying random variable  $Z$ . Let  $\mathcal{W} \subseteq \mathbb{R}^d$  be the space of hypotheses. Suppose that a training sample  $S = (Z_1, Z_2, \dots, Z_n)$  is drawn i.i.d. from  $\mu$  and that a stochastic learning algorithm  $\mathcal{A}$  takes  $S$  as its input and outputs a hypothesis  $W \in \mathcal{W}$  according to some conditional distribution  $P_{W|S}$  mapping  $\mathcal{Z}^n$  to  $\mathcal{W}$ . Let  $\ell : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}^+$  be a loss function, where  $\ell(w, z)$

measures the “unfitness” or “error” of any  $z \in \mathcal{Z}$  with respect to a hypothesis  $w \in \mathcal{W}$ . The population risk, for any  $w \in \mathcal{W}$ , is defined as

$$L_\mu(w) \triangleq \mathbb{E}_{Z \sim \mu}[\ell(w, Z)].$$

The goal of learning is to find a hypothesis  $w$  that minimizes the population risk. But since  $\mu$  is only partially accessible via the sample  $S$ , in practice, we instead turn to the empirical risk, defined as

$$L_S(w) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(w, Z_i).$$

Let  $L_\mu = \mathbb{E}_W [L_\mu(W)]$  be the expected population risk and let  $L_n = \mathbb{E}_{W,S} [L_S(W)]$  be the expected empirical risk. The expected generalization error of the learning algorithm  $\mathcal{A}$  is then defined as

$$\mathcal{E}_\mu(\mathcal{A}) \triangleq L_\mu - L_n = \mathbb{E}_{W,S} [L_\mu(W) - L_S(W)],$$

where the expectation is taken over the joint distribution of  $(S, W)$  (i.e.,  $\mu^n \otimes P_{W|S}$ ).

Furthermore, if we consider the supervised learning setting, then we will let  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  be the domain of the instances, where  $\mathcal{X}$  and  $\mathcal{Y}$  are input and label spaces respectively. A model of interest prescribes a family  $\mathcal{F}$  of predictors,  $\mathcal{F} \subset \mathcal{Y}^\mathcal{X}$ , where each  $f \in \mathcal{F}$  is parameterized by a vector  $w$  in some space  $\mathcal{W}$ . We may write  $f$  as  $f_w$  as needed. Similarly, there is a learning algorithm  $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{W}$ , which takes the training sample  $S$  as the input and outputs a hypothesis  $W \in \mathcal{W}$ , giving rise to a predictor  $f_W \in \mathcal{F}$  that predicts label  $Y$  for input  $X$ .

## 2.3 Other Prerequisite Definitions

**Definition 2.3.1** (Subgaussian). *A random variable  $X$  is  $\sigma$ -subgaussian if for any  $t$ ,  $\log \mathbb{E} \exp(\lambda(X - \mathbb{E}X)) \leq t^2 \sigma^2 / 2$ .*

**Definition 2.3.2** (Wasserstein Distance). *Let  $d(\cdot, \cdot)$  be a metric and let  $P$  and  $Q$  be probability measures on  $\mathcal{X}$ . Denote  $\Gamma(P, Q)$  as the set of all couplings of  $P$  and  $Q$  (i.e. the set of all joint distributions on  $\mathcal{X} \times \mathcal{X}$  with two marginals being  $P$  and  $Q$ ), then the Wasserstein Distance of order one between  $P$  and  $Q$  is defined as  $\mathbb{W}(P, Q) \triangleq \inf_{\gamma \in \Gamma(P, Q)} \int_{\mathcal{X} \times \mathcal{X}} d(x, x') d\gamma(x, x')$ .*

**Definition 2.3.3** (Total Variation). *The total variation between two probability measures  $P$  and  $Q$  is  $\text{TV}(P, Q) \triangleq \sup_E |P(E) - Q(E)|$ , where the supremum is over all measurable set  $E$ .*

**Remark 2.3.1.** Note that the total variation equals to the Wasserstein distance under the discrete metric (or Hamming distortion)  $d(x, x') = \mathbb{1}_{x \neq x'}$  where  $\mathbb{1}$  is the indicator function [44, Theorem 6.15].

**Definition 2.3.4** (Disintegrated Mutual Information [151]). Let  $X, Y$  and  $Z$  be random variables and  $z$  be a realization of  $Z$ . The disintegrated mutual information of  $X$  and  $Y$  given  $Z = z$  is  $I^z(X; Y) \triangleq D_{\text{KL}}(P_{X,Y|Z=z} || P_{X|Z=z} P_{Y|Z=z})$ .

**Remark 2.3.2.** Note that the conditional mutual information  $I(X; Y|Z) = \mathbb{E}_Z I^Z(X; Y)$ .

**Definition 2.3.5** (Lautum Information [162]). The lautum information between  $X$  and  $Y$  is defined as  $L(X; Y) \triangleq D_{\text{KL}}(P_X P_Y || P_{XY})$ .

## 2.4 Technical Lemmas

The key quantity in the most information-theoretic generalization bounds is the mutual information between algorithm's input and output. Specifically, the core technique behind these bounds is the well-known Donsker-Varadhan representation of KL divergence [170, Theorem 3.5].

**Lemma 2.4.1** (Donsker and Varadhan's variational formula). Let  $Q, P$  be probability measures on  $\Theta$ , for any bounded measurable function  $f : \Theta \rightarrow \mathbb{R}$ , we have  $D_{\text{KL}}(Q || P) = \sup_f \mathbb{E}_{\theta \sim Q} [f(\theta)] - \log \mathbb{E}_{\theta \sim P} [\exp f(\theta)]$ .

**Lemma 2.4.2** (Hoeffding's Lemma [94]). Let  $X \in [a, b]$  be a bounded random variable with mean  $\mu$ . Then, for all  $t \in \mathbb{R}$ , we have  $\mathbb{E} [e^{tX}] \leq e^{t\mu + \frac{t^2(b-a)^2}{8}}$ .

**Lemma 2.4.3** (Popoviciu's inequality [171]). Let  $M$  and  $m$  be upper and lower bounds on the values of any random variable  $X$ , then  $\text{Var}(X) \leq \frac{(M-m)^2}{4}$ .

The following lemma is from [140, Lemma 2.8], we provide a self-contained proof.

**Lemma 2.4.4.** Let  $h(x) = \frac{e^x - x - 1}{x^2}$  be the Bernstein function. If a random variable  $X$  satisfies  $\mathbb{E}[X] = 0$  and  $X \leq b$ , then  $\mathbb{E}[e^X] \leq e^{h(b)\mathbb{E}[X^2]}$ .

*Proof.* It's easy to verify that  $h(x)$  is an increasing function for  $x > 0$ . Thus,  $h(x) \leq h(b)$  for  $x \leq b$ . Then,

$$e^x = x + 1 + x^2 h(x) \leq x + 1 + x^2 h(b).$$



For the bounded random variable  $X$  with zero mean, we have

$$\mathbb{E} [e^X] \leq \mathbb{E} [X] + 1 + \mathbb{E} [X^2 h(b)] \leq e^{h(b)\mathbb{E}[X^2]}.$$

The last inequality is by  $e^x \geq x + 1$ . This completes the proof.  $\square$

The following result is known in the previous work [232].

**Lemma 2.4.5** ([232, Lemma 1]). *If  $g(X', Y')$  is  $R$ -subgaussian under  $P_{X', Y'} = P_X P_Y$ , then*

$$|\mathbb{E}_{X, Y} [g(X, Y)] - \mathbb{E}_{X', Y'} [g(X', Y')]| \leq \sqrt{2R^2 I(X; Y)}.$$

In fact, [232, Lemma 1.], is a special case of the following lemma.

**Lemma 2.4.6.** *Let  $Q$  and  $P$  be probability measures on  $\Theta$ . Let  $\theta' \sim Q$  and  $\theta \sim P$ . If  $g(\theta)$  is  $R$ -subgaussian, then,*

$$|\mathbb{E}_{\theta' \sim Q} [g(\theta')] - \mathbb{E}_{\theta \sim P} [g(\theta)]| \leq \sqrt{2R^2 D_{\text{KL}}(Q||P)}.$$

*Proof.* Let  $f = t \cdot g$  for any  $t \in \mathbb{R}$ , by Lemma 2.4.1, we have

$$\begin{aligned} D_{\text{KL}}(Q||P) &\geq \sup_t \mathbb{E}_{\theta' \sim Q} [tg(\theta')] - \log \mathbb{E}_{\theta \sim P} [\exp t \cdot g(\theta)] \\ &= \sup_t \mathbb{E}_{\theta' \sim Q} [tg(\theta')] - \log \mathbb{E}_{\theta \sim P} [\exp t(g(\theta) - \mathbb{E}_{\theta \sim P} [g(\theta)] + \mathbb{E}_{\theta \sim P} [g(\theta)])] \\ &= \sup_t \mathbb{E}_{\theta' \sim Q} [tg(\theta')] - \mathbb{E}_{\theta \sim P} [tg(\theta)] - \log \mathbb{E}_{\theta \sim P} [\exp t(g(\theta) - \mathbb{E}_{\theta \sim P} [g(\theta)])] \\ &\geq \sup_t t (\mathbb{E}_{\theta' \sim Q} [g(\theta')] - \mathbb{E}_{\theta \sim P} [g(\theta)]) - t^2 R^2 / 2, \end{aligned}$$

where the last inequality is by the subgaussianity of  $g(\theta)$ .

Then consider the case of  $t > 0$  and  $t < 0$  ( $t = 0$  is trivial), by AM–GM inequality (i.e. the arithmetic mean is greater than or equal to the geometric mean), the following is straightforward,

$$|\mathbb{E}_{\theta' \sim Q} [g(\theta')] - \mathbb{E}_{\theta \sim P} [g(\theta)]| \leq \sqrt{2R^2 D_{\text{KL}}(Q||P)}.$$

This completes the proof.  $\square$

Let  $\theta = (X, Y)$ ,  $Q = P_{X, Y}$  and  $P = P_X P_Y$ , Lemma 2.4.5 will be recovered from Lemma 2.4.6.

The following lemma is the Kantorovich–Rubinstein duality of Wasserstein distance [44].

**Lemma 2.4.7** (KR duality). *For any two distributions  $P$  and  $Q$ , we have*

$$\mathbb{W}(P, Q) = \sup_{f \in 1\text{-Lip}(\rho)} \int_{\mathcal{X}} f dP - \int_{\mathcal{X}} f dQ,$$

where the supremum is taken over all 1-Lipschitz functions in the metric  $d$ , i.e.  $|f(x) - f(x')| \leq d(x, x')$  for any  $x, x' \in \mathcal{X}$ .

To connect total variation with KL divergence, we will use Pinsker's inequality [170, Theorem 6.5] and Bretagnolle-Huber inequality [27, Lemma 2.1] in this thesis, for more discussion about these two inequalities, we refer readers to [32].

**Lemma 2.4.8** (Pinsker's inequality).  $\text{TV}(P, Q) \leq \sqrt{\frac{1}{2} \text{D}_{\text{KL}}(P||Q)}$ .

**Lemma 2.4.9** (Bretagnolle-Huber inequality).  $\text{TV}(P, Q) \leq \sqrt{1 - e^{-\text{D}_{\text{KL}}(P||Q)}}$ .

Below is the variational formula (or golden formula) of mutual information.

**Lemma 2.4.10** ([170, Corollary 3.1.]). *For two random variables  $X$  and  $Y$ , we have*

$$I(X; Y) = \inf_P \mathbb{E}_X [\text{D}_{\text{KL}}(Q_{Y|X}||P)],$$

where the infimum is achieved at  $P = Q_Y$ .

Below is the KL divergence between two Gaussian distributions  $p = \mathcal{N}(\mu_p, \Sigma_p)$  and  $q = \mathcal{N}(\mu_q, \Sigma_q)$ , where  $\mu_p, \mu_q \in \mathbb{R}^d$  and  $\Sigma_p, \Sigma_q \in \mathbb{R}^{d \times d}$ .

$$\text{D}_{\text{KL}}(p||q) = \frac{1}{2} \left[ \log \frac{\det(\Sigma_q)}{\det(\Sigma_p)} - d + (\mu_p - \mu_q)^{\text{T}} \Sigma_q^{-1} (\mu_p - \mu_q) + \text{tr} \{ \Sigma_q^{-1} \Sigma_p \} \right]. \quad (2.1)$$

The following results are known, and will be widely mentioned throughout this thesis.

**Lemma 2.4.11** ([232, Theorem 1.]). *Assume the loss  $\ell(w, Z)$  is  $R$ -subgaussian for any  $w \in \mathcal{W}$ . The generalization error of  $\mathcal{A}$  is bounded by*

$$|\mathcal{E}_{\mu}(\mathcal{A})| \leq \sqrt{\frac{2R^2}{n} I(W; S)},$$

*Proof.* To apply Lemma 2.4.5, we let  $X = S$ ,  $Y = W$ , and  $g(s, w) = \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$ . Thus, if  $\ell(w, Z)$  is  $R$ -subgaussian for all  $w \in \mathcal{W}$ , then  $g(S, w)$  is  $R/\sqrt{n}$ -subgaussian due to the i.i.d. assumption on  $Z_i$ 's. This implies that  $g(S', W')$  is also  $R/\sqrt{n}$ -subgaussian, where  $S', W'$  are independent copies of  $S$  and  $W$ , respectively. Therefore, utilizing Lemma 2.4.5 and observing that  $\mathcal{E}_{\mu}(\mathcal{A}) = \mathbb{E}[g(S', W')] - \mathbb{E}[g(S, W)]$  completes the proof.  $\square$

## Chapter 3

# On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications

### 3.1 Overview and Main Contributions

In this chapter, we focus on investigating the generalization of machine learning models trained with SGD. Although we are primarily motivated by the curiosity to understanding neural networks, the results of this chapter in fact apply broadly to any model trained with SGD.

This chapter follows the same construction of the auxiliary weight process in [153] and develops upper bounds of generalization error that extend the work of [153]. Like those in [153], the bounds we obtain can be decomposed into two terms, one measuring the impact of training trajectories (“the trajectory term”) and the other measuring the impact of the flatness of the found solution (“the flatness term<sup>1</sup>”). Having an identical flatness term as that in [153], empirical evidence hints that our bounds have an improved trajectory term. Figure 3.1 shows an experimental comparison of the trajectory term in our bound (Theorem 3.4.2) with that in the bound in [153] (re-stated as Lemma 3.3.1 in this chapter) for two neural network models. The trajectory terms are compared in a deterministic

---

<sup>1</sup>We note this term is correlated with flatness rather than precisely measures the flatness.

setting for two different values of  $\sigma$  (the variance parameter of the noise in the auxiliary weight process, details given in appropriate context and in Appendix A.5.2).

The trajectory term in the bounds of [153] accumulates two terms over training steps: *local gradient sensitivity*, measuring the sensitivity of the SGD gradient signal to weight perturbations, and *gradient dispersion*<sup>2</sup>, measuring the extent to which the gradient signal spreads around its mean. Usually gradient dispersion vanishes with training iterations but local gradient sensitivity does not. Our improvement over [153] is achieved by removing the local sensitivity term and involving the logarithm of a revised gradient dispersion. Although our new definition of gradient dispersion is in general larger than that in [153], as long as the number of iterations is not too small, our gradient dispersion also vanishes with training, which allows our bounds to be tighter than [153].

We also provide an application of our bounds in analyzing the generalization behaviour of linear and two-layer ReLU networks, where we show that the activation state in ReLU networks plays an important role in generalization.

It is remarkable that removing the local gradient sensitivity term makes our bound have a simple closed form (after optimizing the noise variances), much easier to evaluate. We empirically validate the derived bounds, and provide various insights pertaining to the generalization behavior of models trained with SGD. For example, gradient dispersion is seen to reveal a double descent phenomenon with respect to training epochs, where the valley in the double descent curve appears to mark the great divide between the “generalization regime” and the “memorization regime” of training. Furthering from this observation, we also show that it is possible to reduce the memorization effect by dynamically clipping the gradient and reducing its dispersion.

Our bounds also inspire a natural and simple solution to alleviate generalization error. Specifically, we propose a new training scheme, referred to as *Gaussian model perturbation* (GMP), aiming at reducing the flatness term of the bounds. This scheme effectively applies a local “average pooling” to the empirical risk surface prior to SGD, greatly resembling the “max-pooling” approach adopted in AMP [244]. We demonstrate experimentally that GMP achieves a competitive performance with the current art of regularization schemes.

Proofs, additional discussions and experimental results are presented in Appendix A.

---

<sup>2</sup>The quantity is often referred to as gradient variance in the literature, but we prefer “dispersion” to “variance” so as to better comply with the mathematical conventions and avoid possible confusion. Note that “dispersion” is a general term used to describe the extent to which a distribution diverges, and “variance” is one of the common measures of “dispersion”.

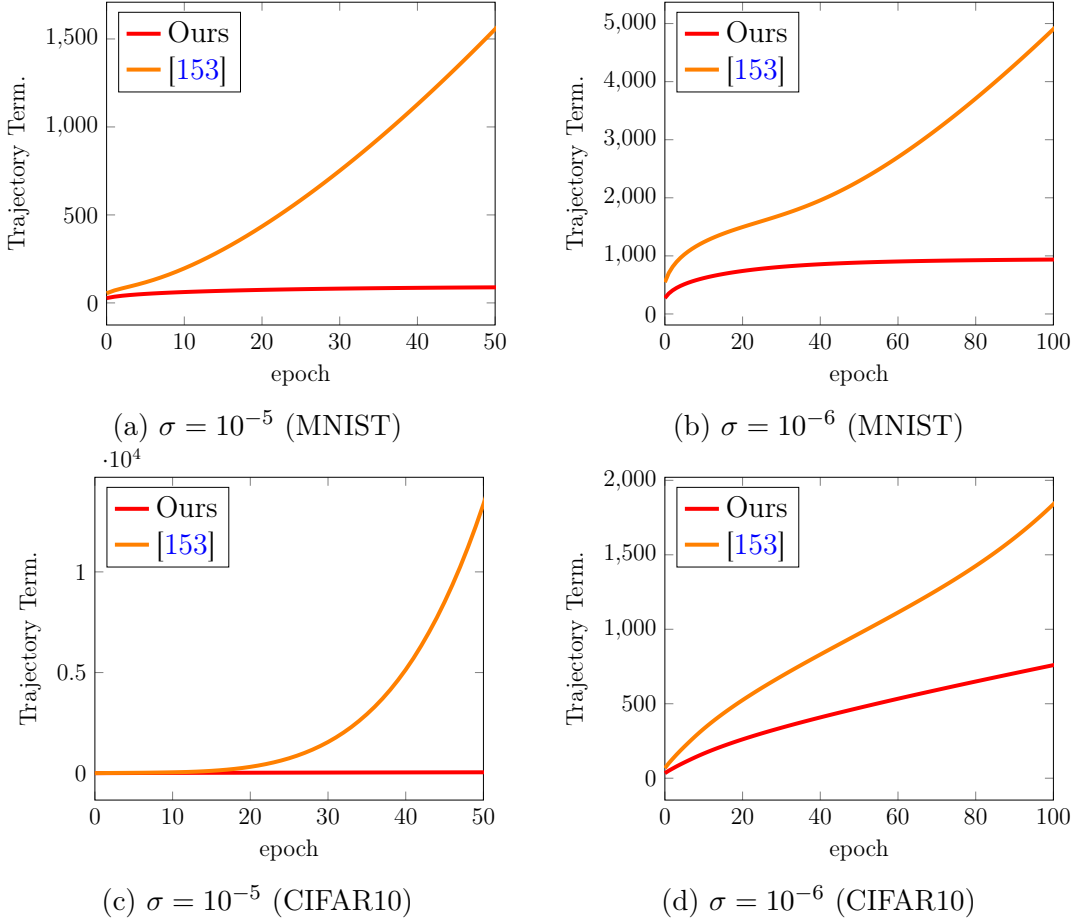


Figure 3.1: Comparison of the trajectory term between our bound (Theorem 3.4.2) and the bound in [153]. (a)(b) MLP trained on MNIST. (c)(d) CNN trained on CIFAR-10.

## 3.2 Other Related Literature

Gradient dispersion is mostly studied from optimization perspectives [23, 182, 104, 221, 53]. Prior to this work, only a few works relate gradient dispersion with the generalization behaviour of the networks. In [153] and [210], gradient dispersion also appears in the generalization bounds. and there are limited studies characterizing the generalization performance and the gradient variance. Particularly, in [102], gradient dispersion is argued to capture a notion of “flatness” of the local minima of the loss landscape, thereby correlating with generalization. Injecting noise in the training process has been proposed in various

regularization schemes [22, 31, 30, 198, 219]. But unlike GMP derived in this chapter, where noise is injected to the model parameters, noise in those schemes is injected either to the training data or to the network activation. In addition, noise can also be implicitly injected during training, for example, by adjusting the momentum terms [230]. The training objective in [228] is similar to our GMP but their NVRM-SGD minimizes a region directly while our GMP considers a trade-off between the empirical loss and the flatness term. Gradient clipping is a common technique for preventing gradient exploding (see, e.g., [142, 169]). This technique is also used in [240] to accelerate training. In this chapter, gradient clipping is used to investigate and control the impact of gradient dispersion on generalization error.

### 3.3 Additional Preliminaries

Throughout this chapter, we take  $\ell$  as a continuous function (adopting the usual notion “surrogate loss” [186]). Additionally, we assume that  $\ell$  is differentiable almost everywhere with respect to both  $w$  and  $z$ . Furthermore we assume that  $\ell(w, Z)$  is  $R$ -subgaussian for any  $w \in \mathcal{W}$ .

**Remark 3.3.1** (Subgaussian Loss). *Note that a bounded loss is guaranteed to be subgaussian, making subgaussian a weaker assumption than bounded loss. However, some popular loss functions such as cross-entropy loss are generally unbounded. Albeit the unboundedness of cross-entropy loss, we believe that it indeed has a subgaussian distribution under the SGD training since the loss remains bounded (and decreasing) during training. In practice, to estimate the variance proxy of the loss when treating it as a subgaussian random variable, one can take the maximum seen value of the loss as its upper bound, and use this bound to obtain an estimate of the variance proxy. Furthermore, in the case of heavy-tailed losses, some additional techniques are required to derive information-theoretic generalization bounds, such as those presented in [134].*

**Stochastic Gradient Descent** We now restrict the learning algorithm  $\mathcal{A}$  to be the mini-batched SGD algorithm for empirical risk minimization. For each training epoch, the dataset  $S$  is randomly split into  $m$  disjoint mini-batches, each having size  $b$ , namely,  $n = mb$ . Based on each batch, one parameter update is performed. Specifically, let  $B_t$  denote the batch used for the  $t^{\text{th}}$  update. Define

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z),$$

namely,  $g(w, B_t)$  is the average gradient computed for the batch  $B_t$  with respect to parameter  $w$ . The rule for the  $t^{\text{th}}$  parameter update is then

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t),$$

where  $\lambda_t$  is the learning rate at the step  $t$ . The initial parameter setting  $W_0$  is assumed to be drawn from the zero-mean spherical Gaussian  $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$  with variance  $\sigma_0^2$  in each dimension. We will assume that the SGD algorithm stops after  $T$  updates and outputs  $W_T$  as the learned model parameter.

Given the training sample  $S$ , let  $\xi$  govern the randomness in the sequence  $(B_1, B_2, \dots, B_T)$  of batches. For the simplicity of notion, we will fix the configuration of  $\xi$ . That is, we will assume a fixed “batching trajectory”, or a fixed way to shuffle the example indices  $\{1, \dots, n\}$  and divide them into  $m$  batches in each epoch. The presented generalization bounds of this chapter can be extended to the case where the batching trajectory is uniformly random (as we set up above). This merely involves averaging over all batching trajectories or taking expectation over  $\xi$ .

**Auxiliary Weight Process** We now associate with the SGD algorithm an auxiliary weight process  $\{\widetilde{W}_t\}$ . Let  $\sigma_1, \sigma_2, \dots, \sigma_T$  be a sequence of positive real numbers. Define

$$\widetilde{W}_0 \triangleq W_0, \quad \text{and} \quad \widetilde{W}_t \triangleq \widetilde{W}_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t, \quad \text{for } t > 0,$$

where  $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}_d)$  is a Gaussian noise. The relationship between this auxiliary weight process  $\{\widetilde{W}_t\}$  and the weight process  $\{W_t\}$  in SGD is shown in the Bayesian network below.

$$\begin{array}{ccccccc}
 & N_1 & N_2 & \cdots & N_{T-1} & N_T \\
 & \downarrow & \downarrow & & \downarrow & \downarrow \\
 \widetilde{W}_0 & \rightarrow \widetilde{W}_1 & \rightarrow \widetilde{W}_2 & \rightarrow \cdots & \rightarrow \widetilde{W}_{T-1} & \rightarrow \widetilde{W}_T \\
 \parallel & \nearrow & \nearrow & \nearrow & \nearrow & \nearrow \\
 W_0 & \rightarrow W_1 & \rightarrow W_2 & \rightarrow \cdots & \rightarrow W_{T-1} & \rightarrow W_T
 \end{array}$$

Let  $\Delta_t = \sum_{\tau=1}^t N_\tau$ . Noting that the weight updates in  $\{\widetilde{W}_t\}$  uses the same gradient signal as that used in  $\{W_t\}$  (which depends on  $W_{t-1}$  not  $\widetilde{W}_{t-1}$ ), so  $\widetilde{W}_t = W_t + \Delta_t$ . Note that this auxiliary process follows the same construction as [153], which we will use to study the generalization error of SGD. To that end, let us define *gradient dispersion* by

$$\mathbb{V}_t(w) \triangleq \mathbb{E} [\|g(w, B_t) - \mathbb{E} [\nabla_w \ell(W, Z)]\|_2^2], \quad (3.1)$$

where the inside expectation is taken over  $(W, Z) \sim \mu \otimes P_{W|Z}$ . For a given sample  $s \in \mathcal{Z}^n$ , define

$$\gamma(w, s) \triangleq \mathbb{E} [L_s(w + \Delta_T) - L_s(w)],$$

where the expectation is taken over  $\Delta_T$  and  $L_s(w)$  is the empirical risk of  $s$  at parameter  $w$ .

In the remainder of the chapter, let  $S'$  denote another sample drawn from  $\mu^{\otimes n}$ , independent of all other random variables. The main generalization bound in [153] is re-stated below.

**Lemma 3.3.1** ([153, Theorem 1.]). *The generalization error of SGD is upper bounded by*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq 2 \sqrt{\frac{2R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E} [\Psi(W_{t-1}) + \tilde{\mathbb{V}}_t(W_{t-1})] + |\mathbb{E} [\gamma(W_T, S) - \gamma(W_T, S')]|},$$

where  $\Psi(w_{t-1}) \triangleq \mathbb{E} [\|\mathbb{E} [\nabla_w \ell(w_{t-1}, Z)] - \mathbb{E} [\nabla_w \ell(w_{t-1} + \zeta, Z)]\|_2^2]$ ,  $\zeta \sim \mathcal{N}(0, \sum_{i=1}^{t-1} \sigma_i^2 \mathbf{I}_d)$  and  $\tilde{\mathbb{V}}_t(w) \triangleq \mathbb{E} [\|g(w, B_t) - \mathbb{E} [\nabla_w \ell(w, Z)]\|_2^2]$ .

The term  $\Psi(w_{t-1})$  in the bound is referred to as “local gradient sensitivity” in [153]. Note that the inside expectation of  $\tilde{\mathbb{V}}_t(w)$  is taken over  $Z \sim \mu$  instead of  $(W, Z) \sim \mu \otimes P_{W|Z}$ . Thus,  $\tilde{\mathbb{V}}_t(w)$  is in general not worse than our gradient dispersion  $\mathbb{V}_t(w)$  for a fixed  $w$  (see Figure A.1 (d) in Appendix A.5.2). However, the difference between these two terms is very small when  $W$  is close to local minima due to the tiny gradient norm. In addition, the definition of our gradient dispersion is implicitly used in a bound in Section 5.2 of [153], which can be regarded as a weaker version of Eq. (3.3) in our Theorem 3.4.2.

## 3.4 New Generalization Bounds for SGD

Removal of the local sensitivity term  $\Psi(w_{t-1})$  requires invoking a special instance of the HWI inequality [174, Lemma 3.4.2], which we first state.

**Lemma 3.4.1.** *Let  $X$  and  $Y$  be two random vectors in  $\mathbb{R}^d$ , and let  $N \sim \mathcal{N}(0, \mathbf{I}_d)$  be independent of  $(X, Y)$ . Then, for any  $t, t' > 0$ ,  $\text{D}_{\text{KL}}(P_{X+\sqrt{t}N} \| P_{Y+\sqrt{t'}N}) \leq \frac{1}{2t'} \mathbb{E} [\|X - Y\|_2^2] + \frac{d}{2} (\log \frac{t'}{t} + \frac{t}{t'} - 1)$ .*



Note that [153] also makes use of a similar lemma in their revision. However, the development in [153] requires constructing a ghost auxiliary weight process in which an independent noise perturbation  $\Delta'_t$  is introduced. Due to the mismatch between  $\Delta_t$  and  $\Delta'_t$ , the local gradient sensitivity term  $\Psi(W_{t-1})$  appears in their bound. In this chapter, we waive this  $\Delta'_t$  by using the following lemma.

**Lemma 3.4.2.** *Let random variables  $X, Y$  and  $\Delta$  be independent of  $N \sim \mathcal{N}(0, \mathbf{I}_d)$ . Then for any  $\sigma > 0$ , any  $\mathbb{R}^d$ -valued function  $f$ , and any random variable  $\Omega \in \mathbb{R}^d$  that is a function of  $Y$ , we have*

$$I(f(Y + \Delta, X) + \sigma N; X|Y) \leq \frac{d}{2} \mathbb{E} \left[ \log \left( \frac{\mathbb{E} [\|f(Y + \Delta, X) - \Omega\|^2]}{d\sigma^2} + 1 \right) \right].$$

Note that the outside expectation is taken over  $Y$  and the inside expectation is taken over  $(X, \Delta)$ . Then, exploiting Lemma 3.4.2 by letting  $X = S$ ,  $Y = \widetilde{W}_t$  and  $\Omega = \mathbb{E}[g(\widetilde{w}_t - \Delta_t, Z)] \triangleq \mathbb{E}[\nabla \ell(\widetilde{w}_t - \Delta_t, Z)]$  will enable us to have the bound below.

**Theorem 3.4.1.** *The generalization error of SGD is upper bounded by*

$$\sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \mathbb{E} \left[ \log \left( \frac{\lambda_t^2 \mathbb{E} [\|g(W_{t-1}, B_t) - \mathbb{E}[g(W_{t-1}, Z)]\|^2]}{d\sigma_t^2} + 1 \right) \right]} + |\mathbb{E}[\gamma(W_T, S) - \gamma(W_T, S')]|.$$

*Proof of Theorem 3.4.1.* We first follow the similar decomposition of the expected generalization error and apply Lemma 2.4.11 as in [153]. Let  $\mathcal{E}_\mu(\mathcal{A})$  be the expected generalization error of SGD, and let  $\mathcal{E}_\mu(\widetilde{\mathcal{A}})$  be the expected generalization error of the corresponding auxiliary weight process. Then,

$$\begin{aligned} |\mathcal{E}_\mu(\mathcal{A})| &= \left| \mathcal{E}_\mu(\widetilde{\mathcal{A}}) + \mathbb{E}_{W_T, \Delta_T} [L_\mu(W_T) - L_\mu(\widetilde{W}_T)] + \mathbb{E}_{W_T, \Delta_T, S} [L_S(\widetilde{W}_T) - L_S(W_T)] \right| \\ &\leq \sqrt{\frac{2R^2}{n} I(\widetilde{W}_T; S)} + |\mathbb{E}_{W_T, S, S'} [\gamma(W_T, S) - \gamma(W_T, S')]|. \end{aligned} \quad (3.2)$$

The remaining task is to bound the mutual information  $I(\widetilde{W}_T; S)$ .

Let  $X = S$ ,  $Y = \widetilde{W}_{t-1}$  and  $\Omega = \mathbb{E}_{\Delta_{t-1}, Z} [g(\widetilde{w}_{t-1} - \Delta_{t-1}, Z)]$ , by applying Lemma 3.4.2, we have

$$\begin{aligned} &I(-\lambda_t g(W_{t-1}, B_t) + N_t; S | \widetilde{W}_{t-1}) \\ &\leq \mathbb{E}_{\widetilde{W}_{t-1}} \left[ \frac{d}{2} \log \left( \frac{\lambda_t^2 \mathbb{E}_{S, \Delta_{t-1}} [\|g(\widetilde{w}_{t-1} - \Delta_{t-1}, B_t) - \mathbb{E}_{\Delta_{t-1}, Z} [g(\widetilde{w}_{t-1} - \Delta_{t-1}, Z)]\|^2 | \widetilde{W}_{t-1} = \widetilde{w}_{t-1}]}{d\sigma^2} + 1 \right) \right]. \end{aligned}$$

By Lemma A.3.1 and putting everything together, we have

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \mathbb{E}_{\widetilde{W}_{t-1}} \left[ \log \left( \frac{\lambda_t^2 \mathbb{E}_{S, \Delta_{t-1}} [\|g(W_{t-1}, B_t) - \mathbb{E}_{Z, \Delta_{t-1}} [g(W_{t-1}, Z)]\|^2]}{d\sigma_t^2} + 1 \right) \right]} + |\mathbb{E} [\gamma(W_T, S) - \gamma(W_T, S')]|.$$

This completes the proof.  $\square$

This bound is tighter than the bound in Lemma 3.3.1. In fact, from Theorem 3.4.1, one can recover Lemma 3.3.1 with a smaller constant factor (see Appendix A.3.2 for more details).

To completely remove  $\Psi(w_{t-1})$  and to obtain a closed form of the optimal bound, we will let  $\Omega = \mathbb{E} [g(W, Z)]$ , then Lemma 3.4.2 gives us a crisp way to have the following upper bound that is independent of the distribution of  $\Delta$ .

**Lemma 3.4.3.** *Let  $G_t = -\lambda_t g(W_{t-1}, B_t)$ . Then,  $I(G_t + N_t; S | \widetilde{W}_{t-1}) \leq \frac{d}{2} \log \left( \frac{\lambda_t^2 \mathbb{E} [\mathbb{V}_t(W_{t-1})]}{d\sigma_t^2} + 1 \right)$ .*

In this lemma, the mutual information  $I(G_t + N_t; S | \widetilde{W}_{t-1})$  roughly indicates the degree by which the SGD's updating signal  $G_t$  (smoothed with noise) depends on the training sample  $S$ , when  $B_t$  is used for computing the gradient. When this dependency is strong (giving rise to a high value of the mutual information), the model conceivably tends to overfit the training sample. This lemma suggests that the strength of this dependency can be upper-bounded by the expected gradient dispersion at the current weight configuration. In our experiments, we will estimate the expected gradient dispersion and validate this intuition.

We are now in a position to state our main theorem.

**Theorem 3.4.2.** *The generalization error of SGD is upper bounded by*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \log \left( \frac{\lambda_t^2 \mathbb{E} [\mathbb{V}_t(W_{t-1})]}{d\sigma_t^2} + 1 \right)} + |\mathbb{E} [\gamma(W_T, S) - \gamma(W_T, S')]|. \quad (3.3)$$

Further, assume  $L_\mu(w_T) \leq \mathbb{E}_\Delta [L_\mu(w_T + \Delta_T)]$ ,  $\ell$  is twice differentiable, and  $\sigma_t^2$  is independent of  $t$ . Denote by  $H_{W_T}$  the Hessian of the loss with respect to  $W_T$  and let  $\text{Tr}(\cdot)$  denote trace. Then

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{3}{2} \left( \sum_{t=1}^T \frac{R^2 \lambda_t^2 T}{n} \mathbb{E} [\mathbb{V}_t(W_{t-1})] \mathbb{E} [\text{Tr} (H_{W_T}(Z))] \right)^{\frac{1}{3}}. \quad (3.4)$$

**Remark 3.4.1.** *With a single draw of  $S$  and under the deterministic setting (i.e., with fixed weight initialization and batching trajectory), removing  $\Psi(W_{t-1})$  makes the bound much tighter as shown in Figure 3.1. This remarkable improvement should come at no surprise, since  $\Psi(W_{t-1})$  monotonically increases with training epochs (noting that  $\Psi(W_{t-1})$  has the cumulative variance  $\sum_{i=1}^{t-1} \sigma_i^2 \mathbf{I}_d$ ) while  $\mathbb{V}_t(W_{t-1})$  appears decreasing (see Figure 3.3a). Figure 3.1 also indicates that if the noise variance  $\sigma_t$  is made small, one expects the gradient dispersion to dominate the trajectory term in Lemma 3.3.1. However, the factor  $1/\sigma_t^2$  will become too large, make the bound loose. More discussions about the stochastic setting with multiple draws of  $S$  are deferred to Appendix A.5.2.*

**Remark 3.4.2.** *The condition  $L_\mu(w_T) \leq \mathbb{E}_{\Delta_T} [L_\mu(w_T + \Delta_T)]$  indicates that the perturbation does not decrease the population risk. This is also assumed in [56] in the derivation of a PAC-Bayesian generalization bound. In addition, this condition leads to the omission of the absolute value function in the generalization error. Specifically, in Eq. (3.4), we bound  $\mathcal{E}_\mu(\mathcal{A})$  directly instead of  $|\mathcal{E}_\mu(\mathcal{A})|$ .*

In the bound of Eq. (3.3), the first term captures the impact of the training trajectory (“trajectory term”), and the second term captures the impact of the final solution, which in fact measures the flatness for the loss landscape at the found solution (“flatness term”). The learning rate in SGD has explicitly appeared in the trajectory term of Eq. (3.4). From the way it appears in the bound, one may be tempted to assert that a small learning rate will improve generalization. This would then contradict some previous observations [100, 224, 84], in which large learning rate will benefit generalization. In addition, the bound also suggests that the batch size has some direct impact on the trajectory term through gradient dispersion. We investigate this by performing experiments with varying learning rates and batch sizes (see Figure A.2 in Appendix A.5.3). As seen in the experiments, increasing the learning rate impacts the trajectory term and the flatness term in opposite ways, i.e. making one increase and the other decrease. A similar (but reverted) behaviour is also observed with batch sizes. This makes the generalization bound in Eq. (3.4), have a rather complex relationship with the settings of learning rate and batch size.

Eq. (3.4) follows from Eq. (3.3) by minimizing the bound over  $\sigma$ . The bound in Eq. (3.4) is thus independent of a choice of  $\sigma_t$ ’s and can be computed easily and efficiently.

To summarize, we remark that these bounds suggest that in order for the model to generalize well, both the trajectory term and the flatness term need to be small—the former involves the learning rate with the gradient dispersion along the training trajectory, whereas the latter depends on the flatness of the empirical risk surface at the found solution.

### 3.5 Application in Linear Networks and Two-Layer ReLU Networks

We now apply Theorem 3.4.2 to two neural network models in a regression setting. Let  $Z = (X, Y)$  with  $X \in \mathbb{R}^{d_0}$  and  $Y \in \mathbb{R}$ . Assume  $\|X\| = 1$ . We use SGD to train a model  $f(W, \cdot) : \mathbb{R}^{d_0} \rightarrow \mathbb{R}$  and define the loss as  $\ell(W, Z) = 1/2(Y - f(W, X))^2$ . The bounds below are both in the form of the product of the flatness term and the trajectory term.

**Theorem 3.5.1** (Linear Networks). *Let the model be a linear network, i.e.  $f(W, X) = W^T X$ , then  $\mathcal{E}_\mu(\mathcal{A}) \leq 3 \left( \sum_{t=1}^T \frac{R^2 \lambda_t^2 T}{4n} \mathbb{E} [\ell(W_{t-1}, Z)] \right)^{\frac{1}{3}}$ .*

Following [8], consider

$$f(W, X) = \frac{1}{\sqrt{m}} \sum_{r=1}^m A_r \text{ReLU}(W_r^T X), \quad (3.5)$$

where  $m$  is the width of the neural network,  $A_r \sim \text{Unif}(\{+1, -1\})$  and  $\text{ReLU}(\cdot)$  is the ReLU activation. We only train the first layer parameters  $W = [W_1, W_2, \dots, W_m] \in \mathbb{R}^{d_0 \times m}$  and fix the second layer parameters  $A = [A_1, A_2, \dots, A_m] \in \mathbb{R}^m$  during training. Then, we have the following generalization bound

**Theorem 3.5.2** (Two-Layer ReLU Networks). *The generalization error of  $f(W, X)$  in Eq. (3.5) is bounded by*

$$\mathcal{E}_\mu(\mathcal{A}) \leq 3 \left( \sum_{r=1}^m \mathbb{E} \left[ \frac{\mathbb{I}_{r,i,T}}{m} \right] \sum_{t=1}^T \frac{R^2 \lambda_t^2 T}{4n} \mathbb{E} \left[ \sum_{r=1}^m \frac{\mathbb{I}_{r,i,t}}{m} \ell(W_{t-1}, Z) \right] \right)^{\frac{1}{3}},$$

where  $\mathbb{I}_{r,i,t} = \mathbb{I}\{W_{t-1,r}^T X_i \geq 0\}$  and  $\mathbb{I}$  is the indicator function.

Compared with Theorem 3.5.1, we notice that in the two-layer ReLU network, the ReLU activation state along the training trajectory plays a key role in the bound of Theorem 3.5.2. Specifically, the weights of the deactivated neurons do not contribute to the bound of Theorem 3.5.2, making the bound not explicitly depend on the model dimension  $d$ . This result also suggests that sparsely activated ReLU networks are expected to generalize better. Despite various empirical evidence pointing to this behaviour (see, e.g., [64]), to the best of our knowledge, this theorem provides the first theoretical justification in this regard.

The comparison between these bounds and the dynamics of generalization gap is deferred to Appendix A.5.4. In the remainder of this chapter, we will study more complex network architectures beyond the linear and two-layer ReLU networks and implications of our bounds in those settings.

## 3.6 Experimental Study

**Bound Verification** We first verify our bound of Eq. (3.4) in Theorem 3.4.2 by training an MLP (with one hidden layer) and an AlexNet [115] on MNIST and CIFAR10 [114], respectively. To simplify estimation, we fix the weight initialization and set  $\sigma_t$  and  $\lambda_t$  to be constants  $\sigma$  and  $\lambda$ , respectively. To compute  $\sum_{t=1}^T \mathbb{E} [\mathbb{V}_t(W_{t-1})]$ , we compute the gradient dispersion as its empirical estimate from a batch, utilizing a PyTorch [164] library BackPack [45]. To compute  $\text{Tr} (\mathbb{E} [\mathbb{H}_{W_T}(Z)])$ , we use the PyHessian library [235] to compute the Hessian.

We perform experiments with varying network width and varying levels of label noise. Specifically, noise level  $\epsilon$  refers to the setting where we replace the labels of  $\epsilon$  fraction of the training and testing instances with random labels. The estimated bound is compared against the true generalization gap, namely, the difference between the training loss and testing loss, and is shown in Figure 3.2.

In Figure 3.2, we see that in all cases the estimated bound follows closely the trend of the true generalization gap. The fact that the bound curve consistently tracks the gap curve under various label noise levels indicates that our bound very well captures the changes of the data distribution. Note that in Figure 3.2 (a) and (b), our bound decays with the increase of the model size, showing a trend as opposite to the bounds obtained in classical learning theory. But such a trend clearly better explains the generalization behaviour of modern neural networks.

**Epoch-wise Double Descent of Gradient Dispersion** We experimentally investigate the impact of gradient dispersion on the training of the neural networks by fixing the learning rate, batch size and weight initialization for the each model (MLP for MNIST, AlexNet for CIFAR10). For each model and various label noise levels, we plot in Figure 3.3 the evolution of the (empirical) gradient dispersion  $\hat{\mathbb{V}}_t(w_{t-1})$ , training accuracy and testing accuracy across training epochs.

An intriguing epoch-wise “double descent” phenomenon is observed, particularly when the labels are noisy. According to the double descent curve, the training may be split

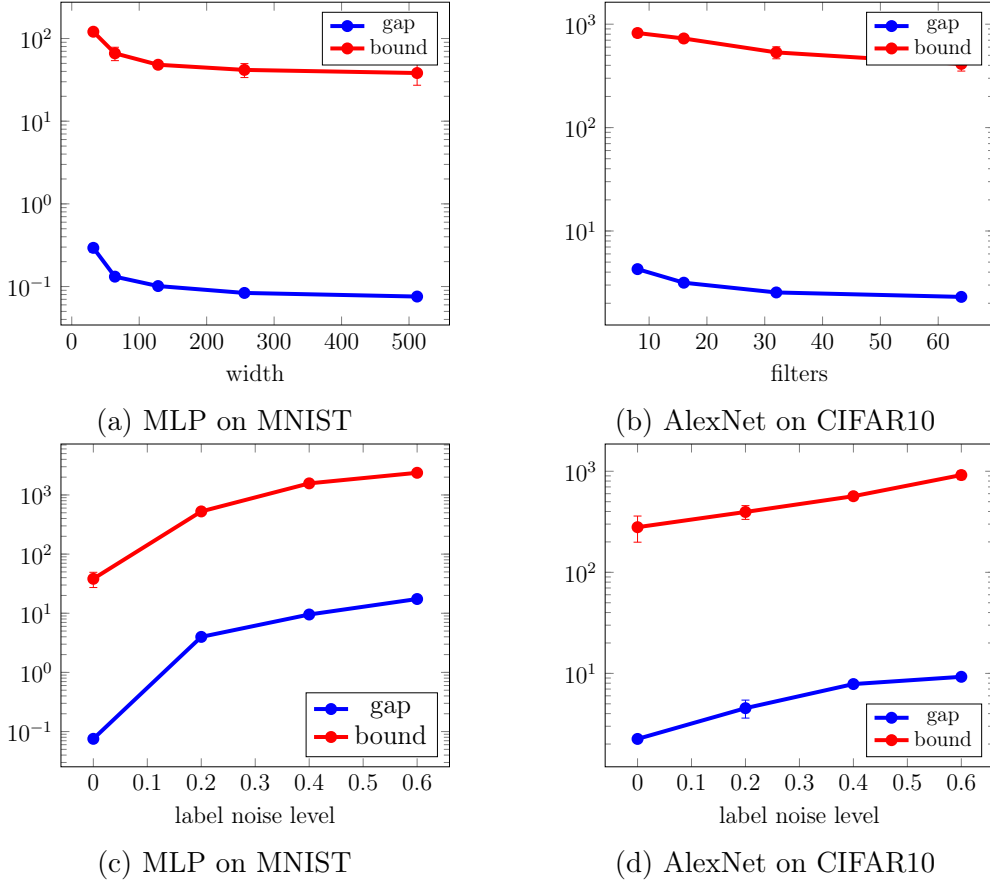


Figure 3.2: Estimated bound and empirical generalization gap (“gap”) as functions of network width ((a) and (b)) and label noise level ((c) and (d)). Note that Y-axis is in log scale.

into three phases (e.g., Figure 3.3 (h)). In the first phase, the gradient dispersion rapidly descends and maintains a very low level. In this phase, both training and test accuracies increase while maintaining a very small generalization gap. This suggests that the network in this phase is extracting useful patterns and generalizes well. In the second phase, the gradient dispersion starts increasing until it reaches a peak value. In this phase, the training and testing accuracies gradually diverge, marking the model entering an overfitting or “memorization” regime – when the data contains the noisy labels, the network mostly tries to memorize the labels in the training set. In the third phase, the gradient dispersion descends again, reaching a low value. In this phase, the model continuously overfits the training data, until the training and testing curves reach their respective maximum and

minimum. It appears that the timing of the three phases depends on the dataset and the label noise level. For simpler data (e.g. MNIST) and cleaner datasets, the first phase may be shorter. This is arguably because in these datasets, extracting useful patterns is relatively easier. Nonetheless, the valley in the double-descent curve appears to mark a “great divide” between generalization and memorization.

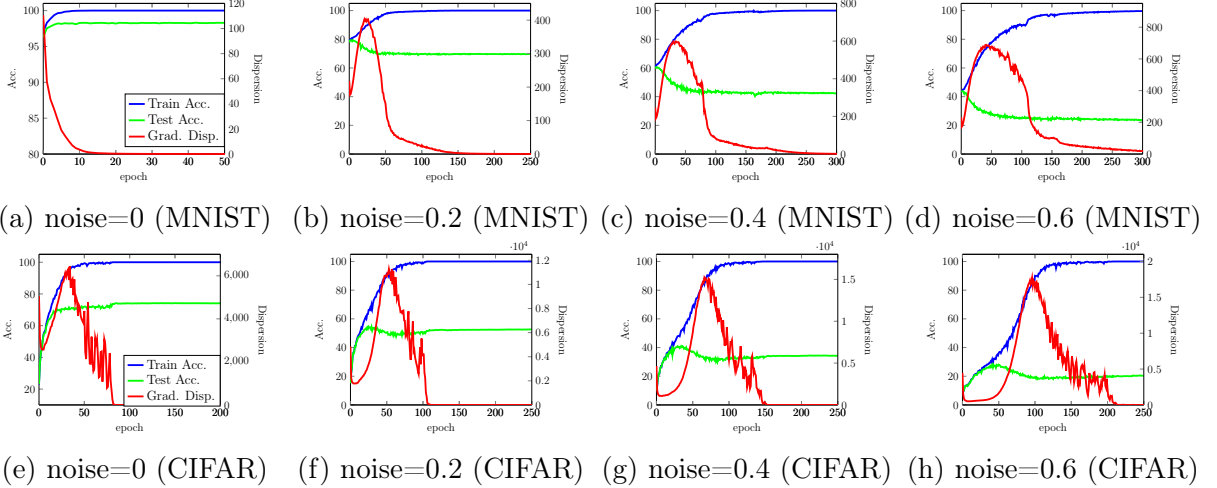


Figure 3.3: Dynamics of gradient dispersion, in relation to training/testing accuracies.

**Dynamic Gradient Clipping** Inspired by our generalization bounds and above observations, one way to reduce the generalization error is to control the trajectory term of the bounds by reducing the gradient dispersion in each training step. Here we investigate a simple scheme that dynamically clips the gradient norm so as to reduce the gradient dispersion. Specifically, whenever the current gradient norm is larger than the gradient norm  $K$  steps earlier, or  $\|g(W_t, B_t)\|_2 > \|g(W_{t-K}, B_{t-K})\|_2$  (i.e., the model is expected to have entered the “memorization” regime), we reduce the norm of the current gradient  $g(W_t, B_t)$  to  $\alpha$  fraction of  $\|g(W_{t-K}, B_{t-K})\|_2$ , for some prescribed value  $\alpha < 1$ .

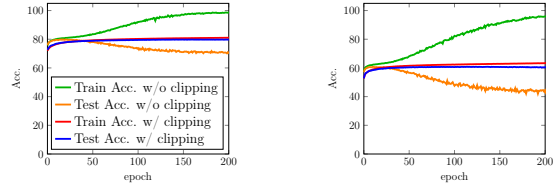


Figure 3.4: Dynamic Gradient Clipping.

The effectiveness of this scheme is best demonstrated when the labels contain noise.

As shown in Figure 3.4 (and Figure A.5 in Appendix A.5.5), dynamic gradient clipping significantly closes the gap between the training accuracy and the testing accuracy. The models trained with this scheme maintain a near-optimal testing accuracy (e.g., about 80% when the label noise level of MNIST is 0.2), without suffering from the severe memorization effect as seen in models trained without this scheme. Further understanding of the double-descent phenomenon of the gradient dispersion may enable more delicate design of such a dynamic clipping scheme and potentially lead to novel and powerful regularization techniques.

### 3.7 A Practical Implication: Gaussian Model Perturbation

The appearance of the flatness term in our generalization bounds suggests that for an empirical risk minimizer  $w^*$  to generalize well, it is necessary that the empirical risk surface at  $w^*$  is flat, or insensitive to a small perturbation of  $w^*$ . This naturally motivates a training scheme using the following regularized loss:

$$\min_w L_s(w) + \rho \mathbb{E}_{\Delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)} [L_s(w + \Delta) - L_s(w)],$$

where  $\rho$  is a hyper-parameter. Replacing the expectation above with its stochastic approximation using  $k$  realizations of  $\Delta$  gives rise to the following optimization problem.

$$\min_w \frac{1}{b} \sum_{z \in B} \left( (1 - \rho) \ell(w, z) + \rho \frac{1}{k} \sum_{i=1}^k (\ell(w + \delta_i, z)) \right).$$

We refer to the SGD training scheme using this loss as *Gaussian model perturbation* or GMP. Notably, GMP requires  $k + 1$  forward passes for every parameter update. Empirical evidence shows that a small  $k$ , for example,  $k = 3$ , already gives competitive performance. Implementing the  $k + 1$  forward passes on parallel processors further reduces the computation load.

We compare GMP with several major regularization schemes in the current art, including Dropout [198], label smoothing (L.S.) [201], Flooding [98], MixUp [239], adversarial training (Adv. Tr.) [67], and AMP [244]. We will also include ERM as the baseline, where no regularization method applied. The compared schemes are evaluated on three popular benchmark image classification datasets SVHN [152], CIFAR-10 and CIFAR-100



[114]. Two representative deep architectures VGG16 [193] and PreActResNet18 [88] are taken as the underlying model. We train the models for 200 epochs by SGD. The learning rate is initialized as 0.1 and divided by 10 after 100 and 150 epochs. For all compared models, the batch size is set to 50 and weight decay is set to  $10^{-4}$ . For GMP, we choose  $\rho = 0.5$  and set the standard deviation of the Gaussian noise  $\Delta$  to 0.03. The value of  $k$  is chosen as 3 and 10 respectively (referred to as GMP<sup>3</sup> and GMP<sup>10</sup>).

| Method                  | SVHN                              | CIFAR-10                          | CIFAR-100                         |
|-------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| ERM                     | 96.86 $\pm$ 0.060                 | 93.68 $\pm$ 0.193                 | 72.16 $\pm$ 0.297                 |
| Dropout                 | 97.04 $\pm$ 0.049                 | 93.78 $\pm$ 0.147                 | 72.28 $\pm$ 0.337                 |
| L. S.                   | 96.93 $\pm$ 0.070                 | 93.71 $\pm$ 0.158                 | 72.51 $\pm$ 0.179                 |
| Flooding                | 96.85 $\pm$ 0.085                 | 93.74 $\pm$ 0.145                 | 72.07 $\pm$ 0.271                 |
| MixUp                   | 96.91 $\pm$ 0.057                 | <b>94.52<math>\pm</math>0.112</b> | 73.19 $\pm$ 0.254                 |
| Adv. Tr.                | 97.06 $\pm$ 0.091                 | 93.51 $\pm$ 0.130                 | 70.88 $\pm$ 0.145                 |
| AMP                     | <b>97.27<math>\pm</math>0.015</b> | 94.35 $\pm$ 0.147                 | 74.40 $\pm$ 0.168                 |
| <b>GMP<sup>3</sup></b>  | <u>97.18<math>\pm</math>0.057</u> | <u>94.33<math>\pm</math>0.094</u> | <u>74.45<math>\pm</math>0.256</u> |
| <b>GMP<sup>10</sup></b> | <u>97.09<math>\pm</math>0.068</u> | <u>94.45<math>\pm</math>0.158</u> | <b>75.09<math>\pm</math>0.285</b> |

Table 3.1: Top-1 classification accuracy acc.(%) of VGG16. We run experiments 10 times and report the mean and the standard deviation of the testing accuracy. Superscript denotes the number of sampled Gaussian noises during training.

cent record given by AMP on SVHN and MixUp on CIFAR-10, respectively [244]. Noting that the key ingredient of AMP, “max-pooling” in the parameter space, greatly resembles regularization term in GMP, which may be seen as “average-pooling” in the same space.

The performances of all compared schemes are given in Table 3.1 (and the results of PreActResNet18 are shown in Appendix A.5.7). For the compared regularization schemes except GMP, we directly report their performances as given in [244]. Table 3.1 demonstrates the effectiveness of GMP. Overall GMP performs comparably to the current art of regularization schemes, although appearing slightly inferior to the most re-

## 3.8 Conclusion

This chapter presents new information-theoretic generalization bounds for models (e.g., linear networks and two-layer ReLU neural networks) trained with SGD. Our bounds naturally point to new and effective regularization schemes. At the same time, our bounds and experimental study reveal interesting phenomena in the SGD training of neural networks. Although our generalization bounds rely on the construction of an auxiliary weight process, which may appear artificial, these guarantees essentially indicate numerous sufficient

conditions for models trained with SGD to perform well. This is further supported by our regularization techniques. There are yet promising directions for further improving these bounds, for example, via exploiting conditional mutual information bounds ([77]), strong data processing inequalities ([210]), and the relationship between the trajectory term and the flatness term. Moreover, extending similar analyses to SGD with momentum is also feasible. However, as also noted in [153], demonstrating the superiority of SGD with momentum over vanilla SGD may be challenging, despite its practical benefits being widely acknowledged.

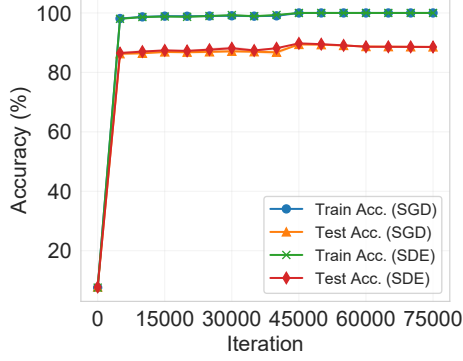
# Chapter 4

## Two Facets of SDE Under an Information-Theoretic Lens: Generalization of SGD via Training Trajectories and via Terminal States

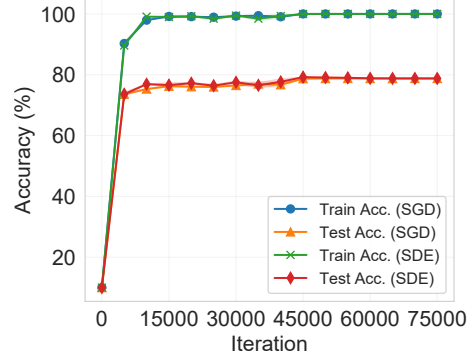
### 4.1 Overview and Main Contributions

Recent research has suggested that the SGD dynamics can be well approximated using stochastic differential equations (SDEs), where the gradient signal in SGD is regarded as the full-batch gradient perturbed with an additive Gaussian noise. Specifically, [137] and [100] model this gradient noise drawn from a Gaussian distribution with a fixed covariance matrix, thereby viewing SGD as performing variational inference. [250, 223, 229, 230] further model the gradient noise as dependent of the current weight parameter and the training data. Modelling SGD in this way provide explanations as to when SGD finds flat minima [250, 229] and sharp minima [251], and inspire some new training techniques [223, 230]. Moreover, [122, 123] and [223] prove that when the learning rate is sufficiently small, the SDE trajectories are theoretically close to those of SGD (cf. Lemma B.1.1). More recently, [125] has demonstrated that the SDE approximation well characterizes the optimization and generalization behavior of SGD without requiring small learning rates.

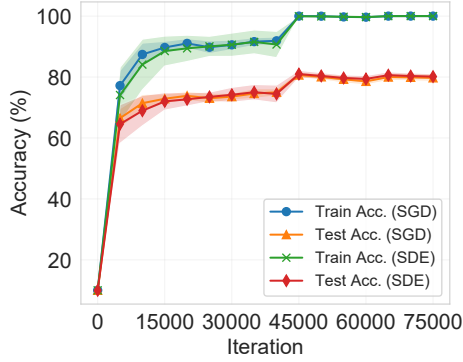
In this chapter, we also empirically verify the consistency between the dynamics of SGD and its associated discrete SDE (cf. Eq. (4.5)). As illustrated in Figure 4.1, the



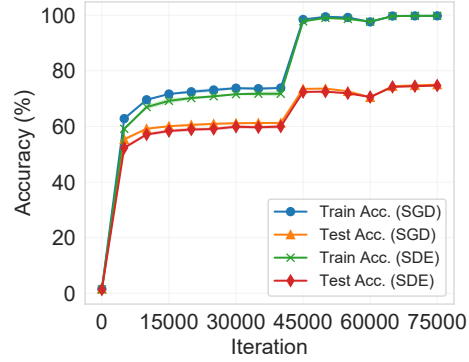
(a) VGG on (small) SVHN



(b) VGG on CIFAR10



(c) ResNet on CIFAR10



(d) ResNet on CIFAR100

Figure 4.1: Performance of VGG-11 and ResNet-18 trained with SGD and SDE. Standard data augmentation techniques are only used in (d).

strong agreement in their performance suggests that, despite the potential presence of non-Gaussian components in the SGD gradient noise, analyzing its SDE through a Gaussian approximation suffices for exploring SGD’s generalization behavior. Furthermore, under the SDE formalism of SGD, SGD becomes an iterative noisy algorithm, on which the aforementioned information-theoretic bounding techniques can directly apply. In particular, we summarize our contributions in this chapter below.

- We obtain a generalization bound (cf. Theorem 4.3.1) in the form of a summation over training steps of a quantity that involves both the sensitivity of the full-batch gradient to the variation of the training set and the covariance of the gradient noise (which makes the SGD gradient deviate from the full-batch gradient). We also give

a tighter bound in Theorem 4.3.2, where the generalization performance of SGD depends on the alignment of the population gradient covariance and the batch gradient covariance. These bounds highlight the significance of (the trace of) the gradient noise covariance in the generalization ability of SGD.

- In addition to the time-dependent trajectories-based bounds, we also provide time-independent (or asymptotic) bounds under some mild assumptions. Specifically, based on previous information-theoretic bounds, we obtain generalization bounds in terms of the KL divergence between the steady-state weight distribution of SGD with respect to a distribution-dependent prior distribution (by Lemma 2.4.11) or data-dependent prior distribution (by Lemma 4.2.1). The former gives us a bound based on the alignment between the weight covariance matrix for each individual local minimum and the weight covariance matrix for the average of local minima (cf. Theorem 4.4.1). Under mild assumptions, we estimate the steady-state weight distribution of SDE (cf. Lemma 4.4.1), leading to a variant of Theorem 4.4.1 (cf. Corollary 4.4.1) and a norm-based bound (cf. Corollary 4.4.2). Additionally, we obtain a stability-based like bound by Lemma 4.2.1 (cf. Theorem 4.4.2), with the notable omission of the Lipschitz constant in other stability-based bounds. Since stability-based bounds often achieve fast decay rates, e.g.,  $\mathcal{O}(1/n)$ , Theorem 4.4.2 provides theoretical advantages compared with other information-theoretic bounds, as it can attain the same rate of decay as the stability-based bound. Comparing to the first family of bounds (i.e., trajectories-based bounds), the second family of bounds directly upper-bound the generalization error via the terminal state, which avoids summing over training steps; these bounds can be tighter when the steady-state estimates are accurate. On the other hand, not relying on the steady-state estimates and the approximating assumptions they base upon is arguably an advantage of the first family.
- We empirically analyze key components within the derived bounds for both algorithms. Our empirical findings reveal that these components for SGD and SDE align remarkably well, further validating the effectiveness of our bounds for assessing the generalization of SGD. Moreover, we provide numerical validation of the presented bounds and demonstrate that our trajectories-based bound is tighter than the result in [213]. Additionally, compared with norm-based bounds, we show that the terminal-state-based bound that integrates the geometric properties of local minima can better characterize generalization.

**Other Related Literature** Recently, [194, 159, 195, 141, 73] challenge the traditional assumption that gradient noise is a Gaussian and argue that the noise is heavy-tailed (e.g., Lévy noise). In contrast, [229] and [125] claim that non-Gaussian noise is not essential to SGD performance, and SDE with Gaussian gradient noise can well characterize the behavior of SGD. They also argue that the empirical evidence shown in [194] relies on a hidden strong assumption that gradient noise is isotropic and each dimension has the same distribution. Other works on SGD and SDE, see [95, 231, 163, 223, 250, 124, 252].

## 4.2 Additional Preliminaries

For a matrix  $A \in \mathbb{R}^{d \times d}$ , we let  $\text{tr}\{A\}$  denote the trace of  $A$  and we use  $\text{tr}\{\log A\}$  to indicate  $\sum_{k=1}^d \log A_{k,k}$ . Furthermore, in the information-theoretic analysis framework, there is a training sample  $S = \{Z_1, Z_2, \dots, Z_n\}_{i=1}^n$  drawn i.i.d. from  $\mu$  and a stochastic learning algorithm  $\mathcal{A}$  takes the training sample  $S$  as its input and outputs a hypothesis  $W \in \mathcal{W}$  according to some conditional distribution  $Q_{W|S}$ . Similar to Chapter 3, throughout this chapter, we take  $\ell$  as a continuous function (adopting the usual notion “surrogate loss”). Additionally, we assume that  $\ell$  is differentiable almost everywhere with respect to  $w$ . In some cases we will assume that  $\ell(w, Z)$  is  $R$ -subgaussian for any  $w \in \mathcal{W}$ . Note that a bounded loss is guaranteed to be subgaussian, for example, if  $\ell(w, Z) \in [0, M]$  then  $R = M/2$ . We will denote  $\ell(w, Z_i)$  by  $\ell_i$  when there is no ambiguity.

**SGD and SDE** At each time step  $t$ , given the current state  $W_{t-1} = w_{t-1}$ , let  $B_t$  be a random subset that is drawn uniformly from  $\{1, 2, \dots, n\}$  and  $|B_t| = b$  is the batch size. Let  $\tilde{G}_t \triangleq \frac{1}{b} \sum_{i \in B_t} \nabla \ell_i$  be the mini-batch gradient. The SGD updating rule with learning rate  $\eta$  is then

$$W_t = w_{t-1} - \eta \tilde{G}_t. \quad (4.1)$$

The full batch gradient is  $G_t \triangleq \frac{1}{n} \sum_{i=1}^n \nabla \ell_i$ . It follows that

$$W_t = w_{t-1} - \eta G_t + \eta V_t, \quad (4.2)$$

where  $V_t \triangleq G_t - \tilde{G}_t$  is the mini-batch *gradient noise*. Since  $\mathbb{E}_{B_t}[V_t] = 0$ ,  $\tilde{G}_t$  is an unbiased estimator of the full batch gradient  $G_t$ . Moreover, the single-draw (i.e.  $b = 1$ ) SGD gradient noise covariance (GNC) and the mini-batch GNC are  $\Sigma_t = \frac{1}{n} \sum_{i=1}^n \nabla \ell_i \nabla \ell_i^{\mathbf{T}} - G_t G_t^{\mathbf{T}}$  and

$C_t = \frac{n-b}{b(n-1)}\Sigma_t$ , respectively. If  $n \gg b$ , then  $C_t = 1/b\Sigma_t$ . Notice that  $\Sigma_t$  (or  $C_t$ ) is state-dependent, i.e. it depends on  $w_{t-1}$ . If  $t$  is not specified, we use  $\Sigma_w$  (or  $C_w$ ) to represent its dependence on  $w$ . In addition, the population GNC at time  $t$  is

$$\Sigma_t^\mu \triangleq \mathbb{E}_Z [\nabla \ell(w_{t-1}, Z) \nabla \ell(w_{t-1}, Z)^\mathbf{T}] - \mathbb{E}_Z [\nabla \ell(w_{t-1}, Z)] \mathbb{E}_Z [\nabla \ell(w_{t-1}, Z)^\mathbf{T}]. \quad (4.3)$$

We assume that the initial parameter  $W_0$  is independent of all other random variables, and SGD stops after  $T$  updates, outputting  $W_T$  as the learned parameter.

We now approximate  $V_t$  up to its second moment, e.g.,  $V_t \sim \mathcal{N}(0, C_t)$ , then we have the following continuous-time evolution, i.e. Itô SDE:

$$d\omega = -\nabla L_S(\omega)dt + [\eta C_\omega]^\frac{1}{2}d\theta_t, \quad (4.4)$$

where  $C_\omega$  is the GNC at  $\omega$  and  $\theta_t$  is a Wiener process. Furthermore, the *Euler-Maruyama* discretization, as the simplest approximation scheme to Itô SDE in Eq. (4.4), is

$$W_t = w_{t-1} - \eta G_t + \eta C_t^{1/2} N_t, \quad (4.5)$$

where  $N_t \sim \mathcal{N}(0, \mathbf{I}_d)$  is the standard Gaussian random variable.

**Validation of SDE** It is important to understand how accurate of SDE in Eq. (4.4) for approximating the SGD process in Eq. (4.1). Previous research, such as [122, 123], has provided theoretical evidence supporting the idea that SDE can approximate SGD in a “weak sense”. That is, the SDE processes closely mimic the original SGD processes, not on an individual sample path basis, but rather in terms of their distributions (see Lemma B.1.1 for a formal result).

Additionally, concerning the validation of the discretization of SDE in Eq. (4.5), [223, Theorem 2] has proved that Eq. (4.5) is *an order 1 strong approximation* to SDE in Eq. (4.4). Moreover, we direct interested readers to the comprehensive investigations carried out by [223, 125], where the authors empirically verify that SGD and Eq. (4.5) can achieve the similar testing performance, suggesting that non-Gaussian noise is not essential to SGD performance. In other words, studying Eq. (4.5) is arguably sufficient to understand generalization properties of SGD. In Figure 4.1, we also empirically verify the approximation of Eq. (4.5), and show that it can effectively capture the behavior of SGD.

**Two Information-Theoretic Bounds** Before we delve into the formal justification, we present below two typical information-theoretic bounds in the literature.

The original version of mutual information based bound is a sample-based MI bound whose main component is the mutual information between the output  $W$  and the entire input sample  $S$ . This result is given in Lemma 2.4.11.

This bound is further improved by a data-dependent prior based bound. Following the setup in [151], let  $J$  be a random subset uniformly drawn from  $\{1, \dots, n\}$  and  $|J| = m > b$ . Let  $S_J = \{Z_i\}_{i \in J}$ . Typically, we choose  $m = n - 1$ , then the following result is known.

**Lemma 4.2.1** ([151, Theorem 2.5]). *Assume the loss  $\ell(w, Z)$  is bounded in  $[0, M]$ , then for any  $P_{W|S_J}$ ,  $\text{Err} \leq \frac{M}{\sqrt{2}} \mathbb{E}_{S,J} \sqrt{\text{D}_{\text{KL}}(Q_{W|S} || P_{W|S_J})}$ .*

Note that  $J$  is drawn before the training starts and is independent of  $\{W_t\}_{t=0}^T$ . We use the subset  $S_J$  to conduct a parallel SGD training process based to obtain a data-dependent prior ( $P_{W|S_J}$ ). When  $m = n - 1$ , we call this prior process the leave-one-out (LOO) prior.

### 4.3 Generalization Bounds Via Full Trajectories

We now discuss the generalization of SGD under the approximation of Eq. (4.5). In particular, we let  $\hat{G}_t = -G_t + C_t^{1/2} N_t$ . We first have the following lemma.

**Lemma 4.3.1.**  $I(\hat{G}_t; S | W_{t-1}) = \mathbb{E}_{W_{t-1}} \left[ \inf_{P_{\hat{G}_t | W_{t-1}}} \mathbb{E}_S^{W_{t-1}} \left[ \text{D}_{\text{KL}}(Q_{\hat{G}_t | S, W_{t-1}} || P_{\hat{G}_t | W_{t-1}}) \right] \right]$ , where the infimum is achieved when the prior distribution  $P_{\hat{G}_t | W_{t-1}} = Q_{\hat{G}_t | W_{t-1}}$  for any  $t$ .

Lemma 4.3.1 suggests that every choice of  $P_{\hat{G}_t | W_{t-1}}$  gives rise to an upper bound of the MI of interest via  $I(\hat{G}_t; S | W_{t-1}) \leq \mathbb{E}_{W_{t-1}} \left[ \mathbb{E}_S^{W_{t-1}} \left[ \text{D}_{\text{KL}}(Q_{\hat{G}_t | S, W_{t-1}} || P_{\hat{G}_t | W_{t-1}}) \right] \right]$ . The closer is  $P_{\hat{G}_t | W_{t-1}}$  to  $Q_{\hat{G}_t | W_{t-1}}$ , the tighter is the bound. As the simplest choice, we will first choose an isotropic Gaussian prior,  $P_{\hat{G}_t | W_{t-1}} = \mathcal{N}(\tilde{g}_t, \sigma_t^2 \mathbf{I}_d)$  (where both  $\tilde{g}_t$  and  $\sigma_t$  are only allowed to depend on  $W_{t-1}$ ), and optimize the KL divergence in Lemma 4.3.1 over  $\sigma_t$  for a fixed  $\tilde{g}_t$ . The following result is obtained.

**Theorem 4.3.1.** *Under the conditions of Lemma 2.4.11 and assume  $C_t$  is a positive-definite matrix. For any  $t \in [T]$ , let  $\tilde{g}_t$  be any constant vector for a given  $w_{t-1}$ , then*

$$\varepsilon_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \mathbb{E}_{W_{t-1}} \left[ d \log \frac{h_1(W_{t-1})}{d} - h_2(W_{t-1}) \right]}, \quad (4.6)$$

where  $h_1(w) = \mathbb{E}_S^w \left[ \|G_t - \tilde{g}_t\|^2 + \text{tr} \{C_t\} \right]$  and  $h_2(w) = \mathbb{E}_S^w [\text{tr} \{\log C_t\}]$ .

Furthermore, if  $\tilde{g}_t = \mathbb{E}_Z [\nabla \ell(w_{t-1}, Z)]$ , then  $h_1(w) = \frac{1}{b} \text{tr} \{\Sigma_t^\mu\}$ .



Notice that  $\tilde{g}_t$  is any reference “gradient” independent of  $S$ , then the first term in  $h_1(W_{t-1})$ ,  $\|G_t - \tilde{g}_t\|^2$ , characterizes the sensitivity of the full-batch gradient to some variation of the training set  $S$ , while the second term in  $h_1(W_{t-1})$ , i.e.  $\text{tr}\{C_t\}$ , reflects the gradient noise magnitude induced by the mini-batch based training. For example, if  $\tilde{g}_t = \mathbb{E}_Z[\nabla\ell(w_{t-1}, Z)]$ , then  $\mathbb{E}_S^{w_{t-1}}[\|G_t - \tilde{g}_t\|^2]$  is the variance of the gradient sample mean, and such  $\tilde{g}_t$  will eventually convert  $h_1(W_{t-1})$  to the population GNC, namely  $h_1(W_{t-1}) = \frac{1}{b}\text{tr}\{\Sigma_t^\mu\}$ .

Moreover, if we simply let  $\tilde{g}_t = 0$ , then Theorem 4.3.1 indicates that one can control the generalization performance via controlling the gradient norm along the entire training trajectories, e.g., if we further let  $b = 1$ , then  $h_1(W_{t-1}) = \frac{1}{n}\sum_{i=1}^n \|\nabla\ell_i\|^2$ . This is consistent with the existing practice, for example, applying gradient clipping [213, 61] and gradient penalty [99, 13, 197, 61] as regularization techniques to improve generalization.

As a by-product, we recover previous information-theoretic bounds for the Gradient Langevin dynamics (GLD) with noise distribution  $\mathcal{N}(0, \eta^2 \mathbf{I}_d)$  below.

**Corollary 4.3.1.** *If  $C_t = \mathbf{I}_d$ , then  $\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \mathbb{E}_{W_{t-1}} \log \left( \mathbb{E}_S^{W_{t-1}} \|G_t - \tilde{g}_t\|^2 / d + 1 \right)}$ .*

It is easy to verify that the bound in Corollary 4.3.1 is tighter than the bound in [153, Proposition 3.] by  $\log(x+1) \leq x$  and it is also stronger than the bound in [168] due to the fact that we use a state dependent quantity  $\mathbb{E}_S^{w_{t-1}}[\|G_t - \tilde{g}_t\|^2]$  rather than some global Lipschitz constant.

While choosing the isotropic Gaussian prior is common in the GLD or SGLD setting, given that we already know  $C_t$  is an anisotropic covariance, one can select an anisotropic prior to better incorporate the geometric structure in the prior distribution. A natural choice of the covariance is a scaled population GNC, namely  $\tilde{c}_t \Sigma_t^\mu$ , where  $\tilde{c}_t$  is some positive state dependent scaling factor. Let  $\tilde{g}_t = \mathbb{E}_Z[\nabla\ell(w_{t-1}, Z)]$  be the state dependent mean. By optimizing over  $c_t$ , we have the bound below.

**Theorem 4.3.2.** *Under the conditions of Lemma 2.4.11 and assume  $C_t$  and  $\Sigma_t^\mu$  are positive-definite matrices, then  $\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \mathbb{E}_{W_{t-1}, S} \left[ \text{tr} \left\{ \log \frac{\Sigma_t^\mu C_t^{-1}}{b} \right\} \right]}$ .*

**Remark 4.3.1.** *If we let the diagonal element of  $\Sigma_t^\mu$  in dimension  $k$  be  $\alpha_t(k)$  and let the corresponding diagonal element of  $\Sigma_t$  be  $\beta_t(k)$ , and assume  $n \gg b$  (so  $\Sigma_t = bC_t$ ), then  $\text{tr} \left\{ \log(\Sigma_t^\mu C_t^{-1}/b) \right\} = \sum_{k=1}^d \log \frac{\alpha_t(k)}{\beta_t(k)}$ . Thus, Theorem 4.3.2 implies that a favorable alignment between the diagonal values of  $\Sigma_t$  and  $\Sigma_t^\mu$  will positively impact generalization performance. In other words, the perfect alignment of these two matrices indicates that*

SGD is insensitive to the randomness of  $S$ . Recall the key quantity in Lemma 2.4.11,  $I(W; S)$ , which also measures the dependence of  $W$  with the randomness of  $S$ , the term  $\Sigma_t^\mu \Sigma_t^{-1}$  conveys a similar intuition in this context.

Compared with Theorem 4.3.1 under the same choice of  $\tilde{g}_t$ , we notice that the main difference is that the term  $\text{tr} \{\log(\Sigma_t^\mu/b)\}$ , instead of  $d \log(\text{tr} \{\Sigma_t^\mu\}/bd)$ , appears in the bound of Theorem 4.3.2. The following lemma demonstrates that Theorem 4.3.2 is tighter than the bound in Theorem 4.3.1.

**Lemma 4.3.2.** *For any  $t$ , we have  $\text{tr} \left\{ \log \frac{\Sigma_t^\mu}{b} \right\} \leq d \log \frac{\text{tr} \{\Sigma_t^\mu\}}{bd}$ , with the equality holds when all the diagonal elements in  $\Sigma_t^\mu$  have the same value, i.e.  $\alpha_t(1) = \alpha_t(2) = \dots = \alpha_t(d)$ .*

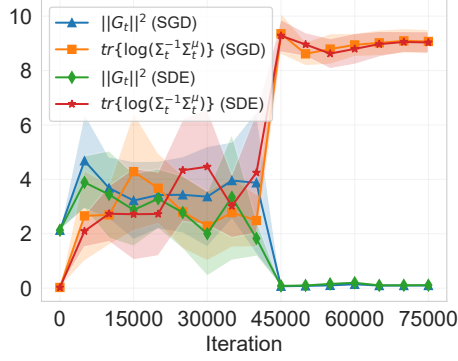
The trajectories-based bounds in Theorem 4.3.1 and Theorem 4.3.2 emphasize the significance of gradient-related information along entire trajectories, including metrics such as gradient norm and gradient covariance alignment, in comprehending the generalization dynamics of understanding the generalization of SGD. In Figure 4.2, we visually show that these key gradient-based measures during SDE training closely mirror the dynamics observed in SGD.

Notably, these trajectory-based information-theoretic bounds are time-dependent, indicating that these bounds may grow with the training iteration number  $T$ , unless the gradient norm becomes negligible at some point during training. While the stability-based bounds for GD/SGD are also time-dependent [80, 16] (in the convex learning case), the learning rate in these bounds helps mitigate the growth of  $T$ . However, the learning rate does not appear in our trajectory-based information-theoretic bounds, making the dependency on  $T$  even worse.

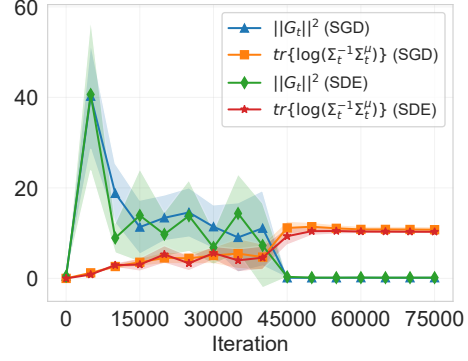
Note that [210] uses the strong data processing inequality to reduce this deficiency, but the bound still increases with  $T$ . To tackle this weakness, we will invoke some asymptotic SDE results on the terminal parameters of the algorithm, which will give us a crisp characterization of the expected generalization gap without decomposing the mutual information.

## 4.4 Generalization Bounds Via Terminal State

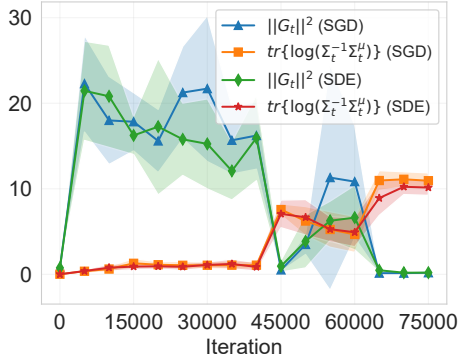
In this section, we directly upper-bound the generalization error by the properties of the terminal state instead of using the full training trajectory information. Particularly, we will first use the stationary distribution of weights at the end of training as  $Q_{W_T|S}$ . To overcome the explicit time-dependence present in the bounds discussed in Section 4.3, one has to



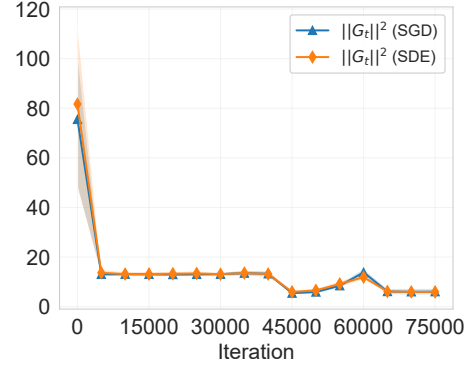
(a) VGG on (small) SVHN



(b) VGG on CIFAR10



(c) ResNet on CIFAR10



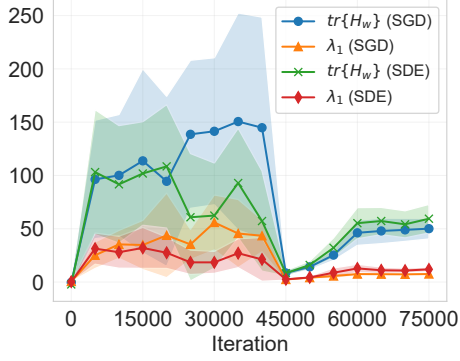
(d) ResNet on CIFAR100

Figure 4.2: Gradient-related quantities of SGD or its discrete SDE approximation. In (d), since per-sample gradient is ill-defined when BatchNormalization is used, we do not track  $tr \{ \log (\Sigma_t^{-1} \Sigma_t^\mu) \}$ .

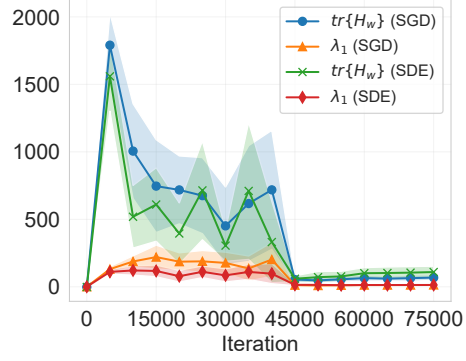
introduce additional assumptions, with these assumptions being the inherent cost. For example, an important approximation used in this section is the quadratic approximation of the loss. Specifically, let  $w_s^*$  be a local minimum for a given training sample  $S = s$ , when  $w$  is close to  $w_s^*$ , we can use a second-order Taylor expansion to approximate the value of the loss at  $w$ ,

$$L_s(w) = L_s(w_s^*) + \frac{1}{2}(w - w_s^*)^\mathbf{T} H_{w_s^*}(w - w_s^*). \quad (4.7)$$

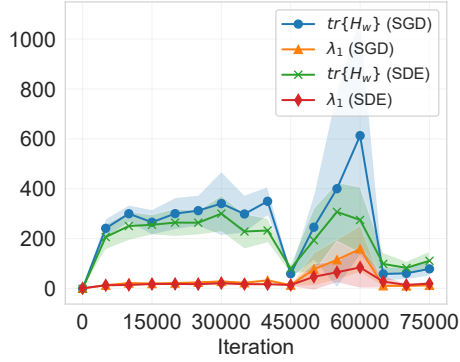
where  $H_{w_s^*}$  is the Hessian matrix of  $s$  at  $w_s^*$ . Note that in this case, when  $w_t \rightarrow w_s^*$ , we have  $G_t = \nabla L_s(w_t) = H_{w_s^*}(w_t - w_s^*)$ . Our remaining analysis assumes the validity of Eq. (4.7).



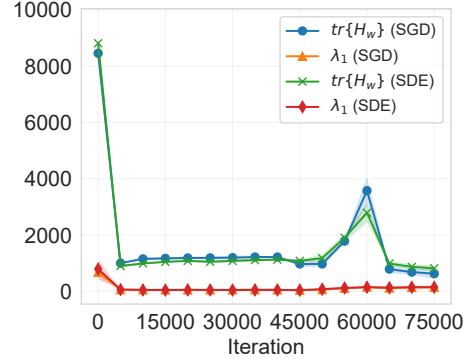
(a) VGG on (small) SVHN



(b) VGG on CIFAR10



(c) ResNet on CIFAR10



(d) ResNet on CIFAR100

Figure 4.3: Hessian-related quantities of SGD or its discrete SDE approximation.

In view of Eq. (4.7), a classical result of [137] shows that the posterior  $Q_{W|s}$  around  $w_s^*$  is a Gaussian distribution  $\mathcal{N}(w_s^*, \Lambda_{w_s^*})$ , where  $\Lambda_{w^*} \triangleq \mathbb{E}[(W - w^*)(W - w^*)^T]$  is the covariance of the stationary distribution (see Appendix B.1.4 for an elaboration). Furthermore, in the context of nonconvex learning, e.g., deep learning, we usually have more than one local minimum  $w_s^*$  for a give  $S = s$ . Hence, it is necessary to take the local minimum itself as a random variable, namely  $W_s^* \sim Q_{W_s^*|s}$ . In this case, we have  $Q_{W|s, w_s^*} = \mathcal{N}(w_s^*, \Lambda_{w_s^*})$  and the posterior distribution  $Q_{W|s} = \mathbb{E}_{W_s^*}^s[\mathcal{N}(W_s^*, \Lambda_{W_s^*})]$  should be a mixture of Gaussian distributions. Moreover, recall that  $I(W_T; S) = \inf_{P_{W_T}} \mathbb{E}_S \text{D}_{\text{KL}}(Q_{W_T|S} || P_{W_T})$  where  $P_{W_T} = Q_{W_T}$  achieves the infimum. Here the oracle prior  $Q_{W_T} = \mathbb{E}_{S, W_s^*}^s[\mathcal{N}(W_s^*, \Lambda_{W_s^*})]$  is also a mixture of Gaussian distributions. From a technique side, given that the KL divergence between two mixture of Gaussian distributions does not have an analytic closed form,

we turn to analyze its upper bound, namely  $\inf_{P_{W_T}} \mathbb{E}_{S, W_S^*} \text{D}_{\text{KL}}(Q_{W_T|S, W_S^*} || P_{W_T})$ . When each  $s$  only has one local minimum,  $I(W; S)$  reaches this upper bound.

We are ready to give the terminal state-dependent bounds.

**Theorem 4.4.1.** *Let  $w_\mu^* = \mathbb{E}[W_S^*]$  and let  $\Lambda_{w_\mu^*} = \mathbb{E}[(W_T - w_\mu^*)(W_T - w_\mu^*)^\text{T}]$  be its corresponding stationary covariance, then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{R}{\sqrt{2n}} \sqrt{\mathbb{E}_{S, W_S^*} \left[ \text{tr} \left\{ \log \left( \Lambda_{W_S^*}^{-1} \Lambda_{w_\mu^*} \right) \right\} \right]}.$$

This result bears resemblance to Theorem 4.3.2 since both involve the alignment between a population covariance matrix and a sample (or batch) covariance matrix.

Note that  $\Lambda_{w_\mu^*} = \mathbb{E}[(W_S^* - w_\mu^*)(W_S^* - w_\mu^*)^\text{T}] + \mathbb{E}[\Lambda_{W_S^*}]$ . By Jensen's inequality, we can move the expectation over  $W_S^*$  inside the logarithmic function. Additionally, if  $\mathbb{E}_{W_S^*} [\Lambda_{W_S^*}^{-1} \mathbb{E}[\Lambda_{W_S^*}]]$  is close to the identity matrix—especially evident in scenarios where each  $s$  has only one minimum, as in convex learning—we obtain the upper bound  $\mathcal{O}\left(\sqrt{\mathbb{E}[\text{d}_M^2(W_S^*, w_\mu^*; \Lambda_{W_S^*})]}/n\right)$ , where  $\text{d}_M(x, y; \Sigma) \triangleq \sqrt{(x - y)^\text{T} \Sigma^{-1} (x - y)}$  is the Mahalanobis distance. Intuitively, this quantity measures the sensitivity of a local minimum to the combined randomness introduced by both the algorithm and the training sample, relative to its local geometry.

In practice, one can estimate  $\Lambda_{w_\mu^*}$  and  $\Lambda_{w_s^*}$  by repeatedly conducting training processes and storing numerous checkpoints at the end of each training run. This is still much easier than estimating  $I(W; S)$  directly. As an alternative strategy, one may leverage the analytical expression available for  $\Lambda_{w_s^*}$ . [137] provides such analysis and give a equation to solve for  $\Lambda_{w_s^*}$ . However, the result in [137] relies on the unrealistic small learning rate, and the GNC in their analysis is regarded as a state-independent covariance matrix. To overcome these limitations, we give the following result under a quadratic approximation of the loss, which is refined from [128, Theorem 1] by using the state-dependent GNC.

**Lemma 4.4.1.** *In the long term limit, we have  $\Lambda_{w^*} H_{w^*} + H_{w^*} \Lambda_{w^*} - \eta H_{w^*} \Lambda_{w^*} H_{w^*} = \eta C_T$ . Moreover, consider the following conditions:*

- (i)  $H_{w^*} \Lambda_{w^*} = \Lambda_{w^*} H_{w^*}$ ;
- (ii)  $H_{w^*}^{-1} \Sigma_T = \mathbf{I}_d$ ;
- (iii)  $\frac{2}{\eta} \gg \lambda_1$ , where  $\lambda_1$  is the top-1 eigenvalue of  $H_{w^*}$ .

Then, given (i), we have  $\Lambda_{w^*} = \left[ H_{w^*} \left( \frac{2}{\eta} \mathbf{I}_d - H_{w^*} \right) \right]^{-1} C_T$ ; given (i-ii), we have  $\Lambda_{w^*} = \left( \frac{2}{\eta} \mathbf{I}_d - H_{w^*} \right)^{-1}$ ; given (i-iii), we have  $\Lambda_{w^*} = \frac{\eta}{2b} \mathbf{I}_d$ .

**Remark 4.4.1.** Note that this result does not rely on Eq. (4.5), and it holds for any type of gradient noise  $V_t$  as long as its covariance is  $C_t$ .

Notably, all the conditions in Lemma 4.4.1 are only discussed in the context of the terminal state of SGD training. Regarding the condition (ii), as being widely used in the literature [100, 250, 124, 229, 230, 128], Hessian is proportional to the GNC near local minima when the loss is the negative log likelihood, i.e. cross-entropy loss. To see this, when  $w_t \rightarrow w^*$ , we have  $\Sigma_{w^*} = \frac{1}{n} \sum_{i=1}^n \nabla \ell_i \nabla \ell_i^T - G_t G_t^T \approx \frac{1}{n} \sum_{i=1}^n \nabla \ell_i \nabla \ell_i^T = F_{w^*}$ , where  $F_{w^*}$  is the *Fisher information matrix* (FIM). This approximation is true because gradient noise dominates over gradient mean near local minima. Moreover, FIM is close to the Hessian near local minima with the log-loss [165, Chapter 8], namely,  $F_{w^*} \approx H_{w^*}$ . Let  $n \gg b$ , we have  $H_{w^*} \approx \Sigma_{w^*} = b C_{w^*}$ . Consequently, when  $\Sigma_T$  is sufficiently close to  $\Sigma_{w^*}$ , condition (ii) is satisfied. It's important to note that the debate surrounding  $H_{w^*} \approx F_{w^*}$  arises when the loss function deviates from cross-entropy [252].

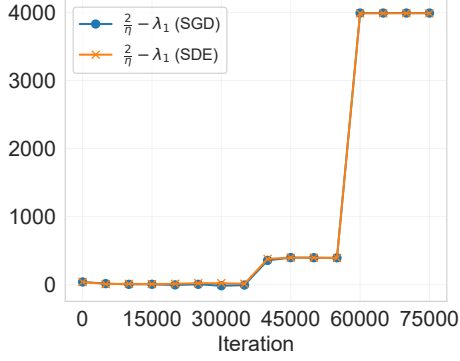
For condition (iii), the initial learning rate is typically set at a high value, and this condition may not be satisfied until the learning rate undergoes decay in the later stages of SGD training. This observation is evident in Figure 4.4a-4.4b, where the condition becomes easily met at the terminal state following the learning rate decay. Moreover, the interplay between  $\frac{2}{\eta}$  and  $\lambda_1$  is extensively explored in the context of the *edge of stability* [224, 38, 10], which suggests that during the training of GD,  $\lambda_1$  approaches  $\frac{2}{\eta}$  and hovers just above it in the “edge of stability” regime. In the context of Theorem 4.4.1, as indicated by Lemma 4.4.1, the diagonal elements of  $\Lambda_{w_s^*}$  tend to be close to zero before reaching the “edge of stability” regime, the bound presented in Theorem 4.4.1 diverges to infinity. This, as a by-product, provides an alternative explanation to the failure mode of  $I(W; S) \rightarrow \infty$  in the deterministic algorithm (e.g., GD with a fixed initialization).

The following results can be obtained by combining Theorem 4.4.1 and Lemma 4.4.1.

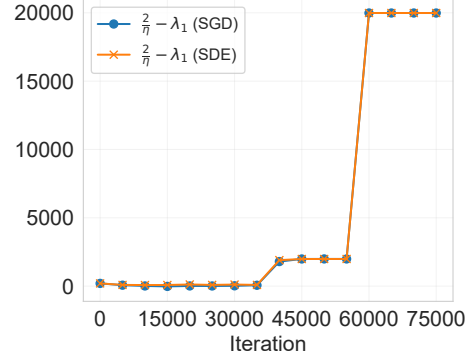
**Corollary 4.4.1.** Under (i,iii) in Lemma 4.4.1, then  $\mathcal{E}_\mu(\mathcal{A}) \leq \frac{R}{\sqrt{n\eta}} \sqrt{\mathbb{E} [\text{tr} \{ \log ([H_{w^*} C_T^{-1}] \Lambda_{w_\mu^*}) \}]}$ .

**Corollary 4.4.2.** Under (i-iii) in Lemma 4.4.1, then  $\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{dR^2}{n} \log \left( \frac{2b}{\eta d} \mathbb{E} [\|W_S^* - w_\mu^*\|^2 + 1] \right)}$ .

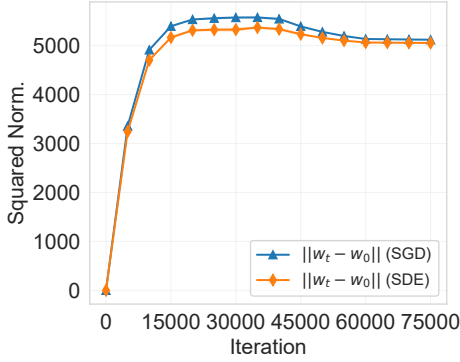
By  $\log(x+1) \leq x$ , the bound in Corollary 4.4.2 is dimension-independent if the weight norm does not grow with  $d$ . Furthermore, the information-theoretic bound becomes a



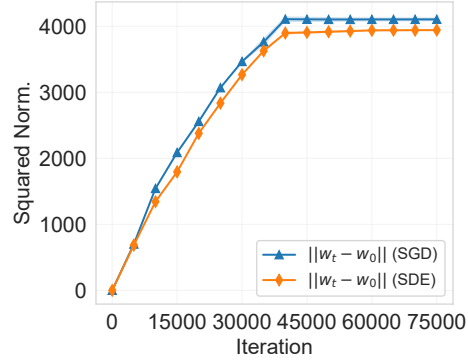
(a) VGG on (small) SVHN



(b) VGG on CIFAR10



(c) VGG on (small) SVHN



(d) VGG on CIFAR10

Figure 4.4: (a-b) The dynamics of  $\eta/2 - \lambda_1$ . Note that learning rate decays by 0.1 at the 40,000<sup>th</sup> and the 60,000<sup>th</sup> iteration. (c-d) The distance of current model parameters from its initialization.

norm-based bound in Corollary 4.4.2, which is widely studied in the generalization literature [14, 155]. In fact,  $w_\mu^*$  can be replaced by any data-independent vector, for example, the initialization,  $w_0$  (see Corollary B.3.1). In this case, the corresponding bound suggests that generalization performance can be characterized by the “distance from initialization”, namely, given that SGD achieves satisfactory performance on the training data, a shorter distance from the initialization tends to yield better generalization. [147] also derived a “distance from initialization” based generalization bound by using Rademacher complexity, and [96] use “distance from initialization” as a regularizer to improve the generalization performance on noisy data.

In the sequel, we use the data-dependent prior bound, namely, Lemma 4.2.1, to derive new results.

**Theorem 4.4.2.** *Let  $P_{W_T|S_J=s_j} = \mathcal{N}(W_{s_j}^*, \Lambda(W_{s_j}^*))$  where  $W_{s_j}^*$  is the local minimum found by the LOO training. Under the same conditions in Lemma 4.2.1 and (i-iii) in Lemma 4.4.1, assuming  $\Lambda(W_{s_j}^*) = \Lambda(W_s^*)$  for a given  $s$ , then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \mathbb{E}_{S,J} \sqrt{\frac{M^2 b}{2\eta} \mathbb{E}_{W_S^*, W_{S_J}^*}^{S,J} \|W_S^* - W_{S_J}^*\|^2}.$$

This bound implies a strong connection between generalization and the algorithmic stability exhibited by SGD. Specifically, if the hypothesis output does not change much (in the squared  $L_2$  distance sense) upon the removal of a single training instance, the algorithm is likely to generalize effectively. In fact,  $\mathbb{E}_{W_S^*, W_{S_J}^*}^{S,J} \|W_S^* - W_{S_J}^*\|^2$  can be regarded as an average version of squared *argument stability* [129]. Moreover, stability-based bounds often demonstrate a fast decay rate in the convex learning cases [80, 16]. It is worth noting that if argument stability achieves the fast rate, e.g.,  $\sup_{s,j} \|w_s^* - w_{s_j}^*\| \leq \mathcal{O}(1/n)$ , then Theorem 4.4.2 can also achieve the same rate. In addition, note that the stability-based bound usually contains a Lipschitz constant, while the bound in Theorem 4.4.2 discards such undesired constant.

Ideally, to estimate the distance of  $\|w_s^* - w_{s_j}^*\|^2$ , one can use the influence function [79, 39, 113], namely  $w_{s_j}^* - w_s^* \approx \frac{1}{n} H_{W_s^*}^{-1} \nabla \ell(w_s^*, z_i)$ , where  $i$  is the instance index that is not selected in  $j$ . However, for deep neural network training, the approximation made by influence function is often erroneous [18]. While this presents a challenge, it motivates further refinement, seeking to enhance the practical application of Theorem 4.4.2 in deep learning.

## 4.5 Empirical Study

In this section, we present some empirical results including tracking training dynamics of SGD and SDE, along with the estimation of several obtained generalization bounds.

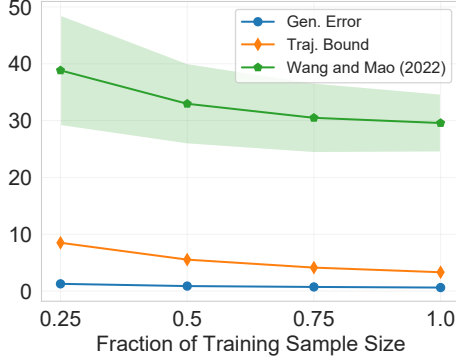
**SGD and SDE Training Dynamics** We implement the SDE training by following the same algorithm given in [223, Algorithm 1]. Our experiments involved training a VGG-11 architecture without BatchNormalization on a subset of SVHN (containing 25k



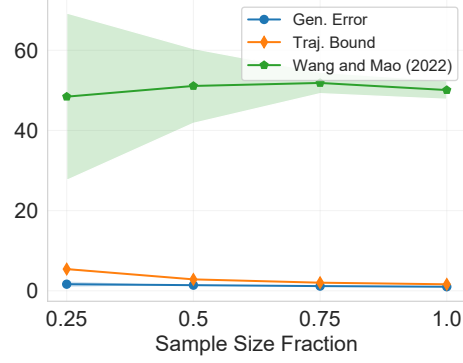
training images) and CIFAR10. Additionally, we trained a ResNet-18 on both CIFAR10 and CIFAR100. Data augmentation is only used in the experiments related to CIFAR100. We ran each experiment for ten different random seed, maintaining a fixed initialization of the model parameters. Further details about the experimental setup can be found in the Appendix. The results are depicted in Figure 4.1. As mentioned earlier, SDE exhibits a performance dynamics akin to that of SGD, reinforcing the similarities in their training behaviors.

**Evolution of Key Quantities for SGD and SDE** We show  $\|G_t\|^2$  and  $\text{tr} \{\log(\Sigma_t^{-1}\Sigma_\mu)\}$  in Figure 4.2. Recognizing the computational challenges for  $\text{tr} \{\log(\Sigma_t^{-1}\Sigma_\mu)\}$ , we opted to draw estimates based on 100 training and 100 testing samples. Notably, both SGD and SDE exhibit similar behaviors in these gradient-based metrics. It is noteworthy that despite the absence of the learning rate in the trajectories-based bounds, we observed that modifications to the learning rate at the 40,000<sup>th</sup> and 60,000<sup>th</sup> steps had discernible effects on these gradient-based quantities. Additionally, in Figure 4.3, we examine the trace of the Hessian and its largest eigenvalue during training, leveraging the PyHessian library [235]. Note that we still use only 100 training data to estimate the Hessian for efficiency. Notice that the Hessian-related quantities of SGD and SDE are nearly perfectly matched in the terminal state of training. Furthermore, Figures 4.4c-4.4d illustrate the "distance to initialization," revealing a consistent trend shared by both SGD and SDE.

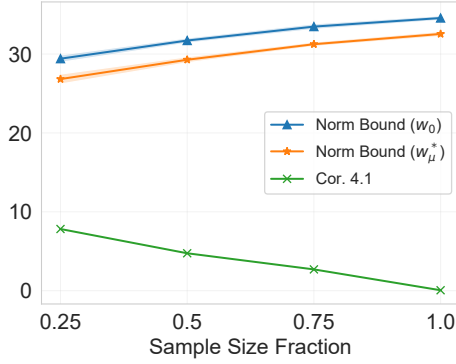
**Bound Comparison** We vary the size of the training sample and empirically estimate several of our bounds in Figure 4.5, with the subgaussian variance proxy  $R$  excluded for simplicity. Thus, the estimated values in Figure 4.5 don't accurately represent the true order of the bounds. Despite the general unbounded nature of cross-entropy loss, common training strategies, such as proper weight initialization, training techniques, and appropriate learning rate selection, ensure that the cross-entropy loss remains bounded in practice. Therefore, it is reasonable to assume subgaussian behavior of the cross-entropy loss under SGD training. In Figure 4.5a-4.5b, we compare our Theorem 4.3.2 with [213, Theorem 2]. Since both bounds incorporate the same  $R$ , the results in Figures 4.5a to 4.5b show that our Theorem 4.3.2 outperforms [213, Theorem 2]. This aligns with expectations (see Appendix B.1.1 for additional discussions), considering that the isotropic Gaussian used in the auxiliary weight process of [213, Theorem 2] is suboptimal, as demonstrated in Lemma 4.3.2. Moreover, Figures 4.5c to 4.5d hint that norm-based bounds Corollary 4.4.2 (and Corollary B.3.1) exhibit growth with  $n$ , which are also observed in [148]. In contrast, Corollary 4.4.1 effectively captures the trend of generalization error, emphasizing the significance of the geometric properties of local minima. Additionally, while trajectories-based



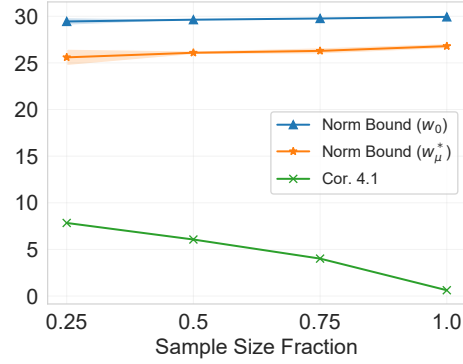
(a) VGG on (small) SVHN



(b) VGG on CIFAR10



(c) VGG on (small) SVHN



(d) VGG on CIFAR10

Figure 4.5: Estimated trajectories-based bound and terminal-state based bound, with  $R$  excluded. Zoomed-in figures of generalization error are given in Figure B.2 in Appendix.

bounds may appear tighter, terminal-state-based bounds seem to have a faster decay rate.

## 4.6 Conclusion

In this chapter, we invoke the SDE approximation of SGD so that information-theoretic generalization bounds are directly applicable to SGD with two opportunities. First, dynamics characterized by SDE enable us to obtain trajectories-based bounds by the step-wise analysis of mutual information. In addition, with some mild assumptions, we also obtain some new bounds based on the terminal state of SGD. While the results in this chapter exhibits certain limitations, such as the requirement of positive definiteness for  $C_t$  in our

trajectories-based bounds, it’s worth noting that recent studies [58, 119, 72, 116] indicate that many parameters in deep neural networks might be dispensable without affecting generalization. This implies that GD/SGD could potentially occur in a subspace of  $\mathbb{R}^d$  termed the “intrinsic dimension”  $d_{\text{int}}$ . Defining  $C_t$  within this invertible subspace, utilizing  $d_{\text{int}}$ , could potentially overcome our current limitations. Theoretical characterization of intrinsic dimension, however, remains an open problem, and further exploration in this direction is poised to significantly improve our work. In addition, there are also some other promising directions for further improving this work, for example, via using other posterior and prior covariance instead of steady-state covariance (e.g., we also give one in Theorem B.5.1 in Appendix), and via extending the similar analytical approach used in this work to other optimizer (e.g., Adam, Adagrad, ...).

# Chapter 5

## Tighter Information-Theoretic Generalization Bounds from Supersamples

### 5.1 Overview and Main Contributions

In this chapter, our development is restricted to the supersample setting of [199], on which we establish novel CMI/MI bounds which are all easy to measure and tighter than the existing bounds in the same setting. Specifically, **1)** we first show that the loss pair used in e-CMI can be replaced by the loss difference, giving rise to a disintegrated CMI bound (cf. Theorem 5.3.1) and an unconditional MI bound (cf. Theorem 5.3.2). Both are tighter than the previous square-root CMI bounds, all within the context of the same supersample construction. In particular, the obtained unconditional MI term can be interpreted as the *achievable rate* over a memoryless channel in communications. We then show that in the interpolating regime (i.e., training error being zero) and under zero-one loss, the generalization error of the learning algorithm can be precisely expressed by the averaged communication rate (cf. Theorem 5.3.3). In other words, we obtain the “tightest bound” of generalization error in this setting. We also establish a novel chained MI bound (cf. Theorem 5.3.4) that is particularly advantageous for continuous and unbounded losses. **2)** Following a symmetric argument for Rademacher process, similar to [245], we explicitly exploit the symmetric structure of expected generalization error by correlating losses with a Rademacher sequence and obtain a novel MI bound involving single losses (cf. Theorem 5.4.1). Using the communication perspective, we show that the MI quantities in the

bound are upper-bounded by the entropy function evaluated at half of the testing error (cf. Theorem 5.4.2). **3)** By correlating losses with a shifted Rademacher sequence, we give novel fast-rate MI bounds of the weighted generalization error (cf. Theorem 5.4.3). **4)** In order to enhance the fast-rate bound in the non-zero training error regime, we extend our analysis by deriving two additional bounds: a variance-based MI bound (cf. Theorem 5.4.4) and a sharpness-based MI bound (cf. Theorem 5.4.5). These novel bounds also incorporate symmetric arguments, as shown in Lemma 5.4.4 and Lemma 5.4.6, respectively. **5)** Experimental results show that our bounds nicely track the generalization dynamics of both linear models and non-linear neural networks, and our fast-rate bounds are tighter than the binary KL bound proposed in [92], the tightest information-theoretic bound known to date for small, non-zero training error. **6)** As a by-product, we also develop a novel Wasserstein distance based bound (cf. Theorem 5.3.5).

Proofs, additional analysis and experimental results are included in Appendix C.

## 5.2 Supersample Setting

The CMI framework for bounding generalization errors is first introduced in [199]. Let  $\tilde{Z} \in \mathcal{Z}^{n \times 2}$  be an  $n \times 2$  matrix, serving as “supersample”, where every entry is drawn i.i.d. from  $\mu$ . For notational convenience, we assume that the columns of  $\tilde{Z}$  are indexed by  $\{0, 1\}$  instead of by  $\{1, 2\}$ . We further denote the  $i$ th row of  $\tilde{Z}$  as  $\tilde{Z}_i$  with entries  $(\tilde{Z}_{i,0}, \tilde{Z}_{i,1})$ . Let  $U = (U_1, U_2, \dots, U_n)^T \sim \text{Unif}(\{0, 1\}^n)$ , independent of  $\tilde{Z}$ , be used to select a training set  $S$  from  $\tilde{Z}$ :  $U_i = 0$  dictates that  $\tilde{Z}_{i,0}$  in  $\tilde{Z}$  be included in the training set  $S$ , and  $\tilde{Z}_{i,1}$  be used for testing;  $U_i = 1$  dictates the opposite. Then, the constructed training sample  $S$  is equivalent to  $\tilde{Z}_U = \{\tilde{Z}_{i,U_i}\}_{i=1}^n$ . Let  $\bar{U}_i = 1 - U_i$ , then the testing sample is  $\tilde{Z}_{\bar{U}} = \{\tilde{Z}_{i,\bar{U}_i}\}_{i=1}^n$ . In addition, let  $L_{i,0} \triangleq \ell(\mathcal{A}(\tilde{Z}_U), \tilde{Z}_{i,0})$  and  $L_{i,1}$  defined similarly. We use  $L_i = (L_{i,0}, L_{i,1})$  to denote the loss pair in the  $i$ th row and  $\Delta L_i = L_{i,1} - L_{i,0}$  be the difference in the pair. To avoid clutter, we might use the superscripts  $+$  and  $-$  to respectively replace the subscripts 0 and 1 in our notations, e.g.,  $\tilde{Z}_i^+ = \tilde{Z}_{i,0}$ ,  $\tilde{Z}_i^- = \tilde{Z}_{i,1}$ ,  $L_i^+ = L_{i,0}$  and  $L_i^- = L_{i,1}$ .

## 5.3 Generalization Bounds via Loss Difference

### 5.3.1 Loss-Difference CMI Bound

Using the loss difference, we first present the following square-root CMI bound.

**Theorem 5.3.1.** *Assume that the loss is bounded between  $[0, 1]$ , we have*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}} \sqrt{2I^{\tilde{Z}}(\Delta L_i; U_i)} \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2I(\Delta L_i; U_i | \tilde{Z})}.$$

Noting the Markov chain  $U - W - f_W(\tilde{Z}_i) - L_i - \Delta L_i$  (conditioned on  $\tilde{Z}$ ) and due to the data-processing inequality (DPI), this “loss-difference CMI” (or “ld-CMI”) bound in Theorem 5.3.1 (the second bound) is tighter than the bound in the previous works [199, 77, 82, 92], namely,  $\underbrace{I(\Delta L_i; U_i | \tilde{Z})}_{\text{ld-CMI}} \leq \underbrace{I(L_i; U_i | \tilde{Z})}_{\text{e-CMI}} \leq \underbrace{I(f_W(\tilde{Z}_i); U_i | \tilde{Z})}_{f\text{-CMI}} \leq \underbrace{I(W; U_i | \tilde{Z})}_{\text{CMI}}.$

It is remarkable that the ld-CMI bound can be significantly tighter. To see this, we re-express the loss function  $\ell$  as a function on  $\mathcal{Y}^2 = \mathcal{Y} \times \mathcal{Y}$ , where  $\ell(y, y')$  is the loss value of the predicted label  $y$  with respect to true label  $y'$ . We say that two elements  $(y_1, y'_1)$  and  $(y_2, y'_2)$  in  $\mathcal{Y}^2$  are *loss-equivalent* and write  $(y_1, y'_1) \equiv_\ell (y_2, y'_2)$  if  $\ell(y_1, y'_1) = \ell(y_2, y'_2)$ . It is straight-forward to verify that  $\equiv_\ell$  is an equivalence relation on  $\mathcal{Y}^2$ . Let  $\mathcal{L}$  denote the image of  $\mathcal{Y}^2$  under  $\ell$ . The quotient space  $\mathcal{Y}^2 / \equiv_\ell$ , or the set of equivalence classes modulo  $\equiv_\ell$ , has a one-to-one correspondence with  $\mathcal{L}$ , under which we may identify  $\mathcal{Y}^2 / \equiv_\ell$  with  $\mathcal{L}$ . Furthermore, we say that two loss pairs  $(\ell_A, \ell'_A)$  and  $(\ell_B, \ell'_B)$  in  $\mathcal{L}^2 = \mathcal{L} \times \mathcal{L}$  are *loss-difference-equivalent* and write  $(\ell_A, \ell'_A) \equiv_\Delta (\ell_B, \ell'_B)$  if  $\ell_A - \ell'_A = \ell_B - \ell'_B$ . Then  $\equiv_\Delta$  is likewise an equivalence relation on  $\mathcal{L}^2$ , which induces the quotient space  $\mathcal{L}^2 / \equiv_\Delta$ . Note that  $f_W(\tilde{Z}_i)$  is a random variable on  $\mathcal{Y}^4 = \mathcal{Y}^2 \times \mathcal{Y}^2$  whereas  $\Delta L_i$  is essentially a random variable on  $\mathcal{L}^2 / \equiv_\Delta$ , which can be identified with  $(\mathcal{Y}^2 / \equiv_\ell)^2 / \equiv_\Delta$  under the aforementioned one-to-one correspondence. There can be a significant reduction of space size from  $\mathcal{Y}^4$  to  $(\mathcal{Y}^2 / \equiv_\ell)^2 / \equiv_\Delta$  when  $\mathcal{Y}$  or  $\mathcal{L}$  is large (assuming they are finite, to fix ideas). Thus,  $\Delta L_i$  reveals much less information about  $U_i$  than  $f_W(\tilde{Z}_i)$  does, making the term  $I(\Delta L_i; U_i | \tilde{Z})$  significantly smaller than  $I(f_W(\tilde{Z}_i); U_i | \tilde{Z})$  and suggesting that the ld-CMI bound can be much tighter than the  $f$ -CMI bound. A similar argument can be made comparing the ld-CMI and the e-CMI bounds.

It is noteworthy that the loss-difference CMI bound is easier to compute than the  $f$ -CMI and e-CMI bounds, since  $\Delta L_i$  is a scalar. Interestingly, when regarding  $\Delta L_i$  as a (scaled) one-dimensional projection of  $L_i$  on a particular direction, the term  $I(\Delta L_i; U_i | \tilde{Z})$  shares some similarity with the notion of *Sliced Mutual Information* (SMI) recently proposed in [65, 66]; the difference is that SMI requires averaging over a random direction of projection.



Figure 5.1: Left: channel from  $U_i$  to  $\Delta L_i$ . Right: channel from  $U_i$  to  $L_i^+$ . Zero-one loss assumed.

### 5.3.2 Loss-Difference MI Bound

Under the setting of supersample as above, we can also obtain a generalization bound based on the loss-difference MI without conditioning on the supersample.

**Theorem 5.3.2.** *Assume that  $\ell(\cdot, \cdot) \in [0, 1]$ , then*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2I(\Delta L_i; U_i)}.$$

By the independence of  $U_i$  and  $\tilde{Z}$ ,  $I(\Delta L_i; U_i) \leq I(\Delta L_i; U_i) + I(U_i; \tilde{Z} | \Delta L_i) = I(\Delta L_i; U_i | \tilde{Z})$ . Then the bound in Theorem 5.3.2 is tighter than ld-CMI bound in Theorem 5.3.1, although not directly comparable to the first bound in Theorem 5.3.1.

It is interesting to relate the MI  $I(\Delta L_i; U_i)$  to a communication setting where  $P_{\Delta L_i | U_i}$  specifies a memoryless channel with input  $U_i$  and output  $\Delta L_i$ . Then  $I(\Delta L_i; U_i)$  is the *rate of reliable communication* over this channel achievable with the input distribution  $P_{U_i}$  (which is  $\text{Bern}(\frac{1}{2})$  by the construction of  $U$ ) [189]. Consider the special case where  $\ell(\cdot, \cdot)$  is the *zero-one loss*, i.e.,  $\ell(w, z) = \mathbb{1}_{f_w(x) \neq y}$ . In this case,  $\Delta L_i \in \{-1, 0, 1\}$ , and the channel is shown in Figure 5.1 (left), in which  $\epsilon_i$  and  $\alpha_i$  are transition probabilities as shown on the respective transition edges. In particular, recalling  $\Delta L_i = L_i^- - L_i^+$ , we see that  $\alpha_i$  is the probability that in  $\tilde{Z}_i$  the instance selected from training has the same loss value as that selected for testing, and that  $\epsilon_i$  is the probability that the training instance in  $\tilde{Z}_i$  has a higher loss value than the testing instance. It follows that any *interpolating algorithm*, namely, one that achieves zero training error must have  $\epsilon_i = 0$  for each  $i$ . The following theorem can then be proved.

**Theorem 5.3.3.** *Under zero-one loss and for any interpolating algorithm  $\mathcal{A}$ ,  $I(\Delta L_i; U_i) = (1 - \alpha_i) \ln 2$  nats for each  $i$ , and  $|\mathcal{E}_\mu(\mathcal{A})| = L_\mu = \sum_{i=1}^n \frac{I(\Delta L_i; U_i)}{n \ln 2}$ .*

In this case, the generalization error is exactly determined by the communication rate over the channel in Figure 5.1 (left) averaged over all such channels, making Theorem 5.3.3 the obviously the “tightest bound” of generalization error in the “interpolating regime”. It is of course also tighter than the interpolating bound in [92], which may be alternatively seen from  $I(\Delta L_i; U_i) \leq I(L_i; U_i | \tilde{Z})$ . Note that [76] also gives a MI quantity that can determine the generalization error in the interpolating case, although their leave-one-out MI is between an  $n + 1$ -dimensional random variable and an one-dimensional random variable, and its corresponding bound is established without exploiting the communication perspective.

Furthermore, it is possible to establish further tightened loss-difference MI bounds for more general loss functions than those required in Theorem 5.3.2. Specifically, the loss function can be unbounded and continuous, as presented in next theorem, where we apply the chaining technique [11, 74, 248, 37] and the obtained bound consists of MI terms between  $U_i$  and the successively quantized versions of  $\Delta L_i$ . To that end, let  $\mathcal{E}_\mu(\mathcal{A})^i(\Delta \ell_i) \triangleq (-1)^{U_i} \Delta \ell_i$  and let  $\Gamma \subseteq \mathbb{R}$  be the range of  $\Delta \ell$ . Then  $\{\mathcal{E}_\mu(\mathcal{A})^i(\Delta \ell_i)\}_{\Delta \ell_i \in \Gamma}$  is a random process<sup>1</sup>, applying the *stochastic chaining* method [248] gives the following chained MI bound.

**Theorem 5.3.4.** *For each  $i \in [n]$ , we assume  $\{\Delta L_{i,k}\}_{k=k_0}^\infty$  is a stochastic chain<sup>1</sup> of  $(\{\mathcal{E}_\mu(\mathcal{A})^i(\Delta \ell_i)\}_{\Delta \ell_i \in \Gamma}, \Delta L_i)$ , then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{1}{n} \sum_{i=1}^n \sum_{k=k_0}^\infty \sqrt{2\mathbb{E}[|\Delta L_{i,k} - \Delta L_{i,k-1}|^2] I(\Delta L_{i,k}; U_i)},$$

where  $\Delta L_{i,k}$  is the  $k$ th level of quantization of  $\Delta L_i$ , the RHS expectation is taken over  $(\Delta L_{i,k}, \Delta L_{i,k-1})$ .

Notice that the bound is expressed as MI terms each involving  $U_i$  and  $\Delta L_{i,k}$ , both being discrete random variables. This has not arose in the previous chained weight-based MI bounds where they either contain the continuous random variable  $S$  [11, 248, 37] or are conditioned on the continuous random variable  $\tilde{Z}$  [74]. Additionally, by the master definition of MI [41, Eq. (8.54)], we know that  $I(\Delta L_i; U_i) = \sup_k I(\Delta L_{i,k}; U_i)$ , and  $I(\Delta L_{i,k}; U_i) \rightarrow I(\Delta L_i; U_i)$  when  $k \rightarrow \infty$ .

For bounded loss, the diameter  $\text{diam}(\Gamma)$  is finite, we can use hierarchical partitions as in [11] to construct a deterministic sequence of  $\{\Delta L_{i,k}\}_{k=k_0}^\infty$ . This is deferred to Corollary C.2.1 in Appendix.

---

<sup>1</sup>Some prerequisite definitions of the chaining technique (such as *stochastic chain*, *separable process* and *subgaussian process*) are give in the Appendix C.1.



### 5.3.3 Loss-Difference Bound Beyond CMI and MI

It is possible to develop generalization bounds based on the loss differences in the super-sample using distances or divergences beyond the information-theoretic measures. Here we present such a bound based on Wasserstein distance. As investigated in the previous literature [180], Wasserstein distance usually gives a tighter bounds than the mutual information.

**Theorem 5.3.5.** *Assume that  $\ell(\cdot, \cdot) \in [0, 1]$ , then*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{U_i} [\mathbb{W}(P_{\Delta L_i|U_i}, P_{\Delta L_i})].$$

Unlike the results in [180], here we do not require the loss to be Lipschitz continuous.

## 5.4 Generalization Bounds via Correlating with Rademacher Sequence

We have so far obtained tighter square-root MI bounds based on the information measures (and their variants) between the loss difference  $\Delta L_i$  and the mask variable  $U_i$ . However, the loss difference may not be used to obtain the fast-rate generalization bound where the square root function is removed [68, 90, 92]. This is because deriving the fast-rate bound usually relies on a weighted generalization error, for which one loses the center-symmetric structure of the standard generalization error. Specifically, knowing  $\Delta L_i$  and  $U_i$  is sufficient to determine the generalization error at  $i$ th position by  $(-1)^{U_i} \Delta L_i$ . However, for the weighted generalization error at  $i$ th row defined by  $E_{C_1}^i = L_{i, \bar{U}_i} - C_1 L_{i, U_i}$  (for some constant  $C_1 > 0$ ), having  $U_i$  and a weighted loss difference  $\Delta_{C_1} L_i = L_i^- - C_1 L_i^+$ , does not allow its recovering from  $(-1)^{U_i} \Delta_{C_1} L_i$  since  $L_i^+ - C_1 L_i^- \neq C_1 L_i^+ - L_i^-$  in general. Indeed, knowing both  $L_i^- - C_1 L_i^+$  and  $L_i^+ - C_1 L_i^-$  requires knowing  $L_i$ . Then in order to obtain fast-rate bounds, we need to give up the loss difference and return to the original e-CMI as in [92].

Therefore, if we still want to use a MI between two one-dimensional random variables to bound the error, we need to find another trick. This motivates us to use a Rademacher viewpoint to derive the CMI bounds. Before we handle the fast-rate CMI bound for the weighted generalization error, we again consider the standard generalization error.

### 5.4.1 Single-Loss MI Bounds

Although the CMI setting, particularly its construction of the “ghost sample”, is conceptually related to the Rademacher complexity [15], the information-theoretic generalization bounds in previous literature do not explicitly exploit this connection. Fortunately, both information-theoretic bounds [151, 89, 90, 92] and the Rademacher viewpoint [108, 233] are shown connected to the PAC-Bayes bounds, we thus derive a variant of e-CMI bound by invoking a similar symmetric argument with [245, 233].

We first note the following lemma.

**Lemma 5.4.1.** *The expected generalization error  $\mathcal{E}_\mu(\mathcal{A}) = \frac{2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \varepsilon_i} [\varepsilon_i L_i^+]$ , where  $\varepsilon_i = (-1)^{\bar{U}_i}$ .*

Note that  $\varepsilon_{1:n} = \{\varepsilon_i\}_{i=1}^n$  is a sequence of Rademacher random variables, and the lemma suggests that  $\mathcal{E}_\mu(\mathcal{A}) = 2\mathbb{E}_{\tilde{Z}_{1:n}^+} \mathbb{E}_{\varepsilon_{1:n}} \mathbb{E}_{L_{1:n}^+ | \varepsilon_{1:n}, \tilde{Z}_{1:n}^+} [\frac{1}{n} \sum_{i=1}^n \varepsilon_i L_i^+]$ , where  $\tilde{Z}_{1:n}^+ = \{\tilde{Z}_i^+\}_{i=1}^n$  and  $L_{1:n}^+ = \{L_i^+\}_{i=1}^n$ . Then, recall that the Rademacher complexity is defined as  $\mathfrak{R}_n(\mathcal{W}) \triangleq \mathbb{E}_S \mathbb{E}_{\varepsilon_{1:n}} [\sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell(w, Z_i)]$  [15]. Notably, the expected generalization error can be viewed, up to a scale factor 2, as an “average” version of the Rademacher complexity. While  $\mathcal{E}_\mu(\mathcal{A})$  considers the average correlation between the loss sequence and the Rademacher sequence, the Rademacher complexity measures the worst such correlation. Thus,  $\mathcal{E}_\mu(\mathcal{A}) \leq 2\mathfrak{R}_n(\mathcal{W})$ .

Based on Lemma 5.4.1, we have the following bound.

**Theorem 5.4.1.** *Assume  $\ell(\cdot, \cdot) \in [0, 1]$ , we have*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{2}{n} \sum_{i=1}^n \sqrt{2I(L_i^+; U_i)} \leq \frac{2}{n} \sum_{i=1}^n \sqrt{2I(f_W(X_i^+); U_i | \tilde{Z})}.$$

The variable  $U_i$  in the above MI/CMI terms can obviously be replaced by  $\varepsilon_i$ . Thus the theorem can be interpreted as using a different notion of “average correlation”, namely mutual information, between losses (or predictions) and Rademacher noises to bound the original notion of average correlation (as stated in Lemma 5.4.1 and discussed earlier).

This bound may not be directly comparable to others due to the undesired constant of 2 outside of the square root function in the bound. We will soon see that  $I(L_i^+; U_i)$  based bound will be more useful when the square root is removed.

For the zero-one loss, the dependence between  $U_i$  and  $L_i^+$  is characterized by the communication channel given in Figure 5.1 (right). In this case,  $U_i = 0$  indicates  $\tilde{Z}_i^+$  is selected

for training, then  $p_i$  is the error rate on this training instance. Similarly, when  $U_i = 1$ ,  $\tilde{Z}_i^+$  is used for testing, then  $1 - q_i$  is the error rate on this testing instance. In practice, we usually have  $p_i < 1 - p_i$  since  $L_i^+$  is more likely to be zero when  $\tilde{Z}_i^+$  is a training instance, and we may also have  $p_i < 1 - q_i$  since  $L_i^+$  is more likely to be one when  $\tilde{Z}_i^+$  is a testing instance compared with the case when  $\tilde{Z}_i^+$  is used in training. When  $p_i = 0$ , this channel reduces to a  $Z$ -channel [41]. This corresponds to an interpolating algorithm, for which we have the following theorem.

**Theorem 5.4.2.** *For zero-one loss and any interpolating algorithm, we have  $\frac{1}{n} \sum_{i=1}^n I(L_i^+; U_i) \leq H(\frac{L_\mu}{2})$ .*

When the loss is not discrete, we can again obtain a chained MI bound by quantizing the continuous random variable  $L_i^+$ , which is given in Theorem C.3.1 in Appendix C.3.4.

### 5.4.2 Fast-Rate MI Bound

We are now in a position to discuss the weighted generalization error,  $\mathcal{E}_\mu^{C_1}(\mathcal{A}) \triangleq L_\mu - (1 + C_1)L_n$ , where  $C_1$  is a prescribed constant. This notion is important for obtaining the fast-rate PAC-Bayes bounds [33].

To bound this weighted generalization error, similar to Lemma 5.4.1, we have the following symmetry argument.

**Lemma 5.4.2.** *The weighted generalization error can be rewritten as*

$$\mathcal{E}_\mu^{C_1}(\mathcal{A}) = \frac{2 + C_1}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i} [\tilde{\varepsilon}_i L_i^+],$$

where  $\tilde{\varepsilon}_i = (-1)^{\bar{U}_i} - \frac{C_1}{C_1+2}$  is a shifted Rademacher variable with mean  $-\frac{C_1}{C_1+2}$ .

The relationship between Err and Rademacher complexity also likewise extends to that between  $\mathcal{E}_\mu^{C_1}(\mathcal{A})$  and “shifted Rademacher complexity”, which is defined as

$$\tilde{\mathfrak{R}}_n(\mathcal{W}) \triangleq \mathbb{E}_S \mathbb{E}_{\tilde{\varepsilon}_{1:n}} \left[ \sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \tilde{\varepsilon}_i \ell(w, Z_i) \right],$$

namely  $\mathcal{E}_\mu^{C_1}(\mathcal{A}) \leq 2\tilde{\mathfrak{R}}_n(\mathcal{W})$ .

Then, we are ready to present the following bounds.

**Theorem 5.4.3.** *Let  $\ell(\cdot, \cdot) \in [0, 1]$ . There exist  $C_1, C_2 > 0$  such that*

$$L_\mu \leq (1 + C_1)L_n + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n}, \quad (5.1)$$

$$L_\mu \leq L_n + \sum_{i=1}^n \frac{4I(L_i^+; U_i)}{n} + 4\sqrt{\sum_{i=1}^n \frac{L_n I(L_i^+; U_i)}{n}}. \quad (5.2)$$

Furthermore, if  $\mathcal{A}$  is an interpolating algorithm, we have

$$L_\mu \leq \sum_{i=1}^n \frac{2I(L_i^+; U_i)}{n \ln 2}. \quad (5.3)$$

Notice that Eq. (5.2) does not depend on  $C_1, C_2$ , as it is obtained via minimizing the bound in Eq. (5.1) over a region of  $(C_1, C_2)$  in which Eq. (5.1) hold.

Comparing Eq. (5.3) with the interpolating bound in [92, Eq. (12)], the main difference is that their bounds<sup>2</sup> are based on  $I(L_i^+, L_i^-; U_i)$ , instead of  $2I(L_i^+; U_i)$ . This difference could be characterized by the *interaction information* [236], namely  $I(L_i^+; U_i; L_i^-) = I(L_i^+; U_i) - I(L_i^+; U_i | L_i^-) = 2I(L_i^+; U_i) - I(L_i; U_i)$  (where the second equality is by the chain rule of MI), and the value  $I(L_i^+; L_i^-; U_i)$  could be positive, negative and zero. Hence, the interpolating bound could be further improved as below

$$L_\mu \leq \sum_{i=1}^n \frac{\min\{2I(L_i^+; U_i), I(L_i; U_i)\}}{n \ln 2}. \quad (5.4)$$

This bound is strictly non-vacuous since the RHS of Eq. (5.2) is upper-bounded by  $\frac{\sum_{i=1}^n H(U_i)}{n \ln 2} = 1$ . Note that the “tightest bound” of the interpolating algorithm is already obtained in Theorem 5.3.3.

Previous works [199, 92] suggest that the fast-rate bounds for the weighted generalization error are typically useful when the empirical risk is small or even zero, which may restrict their applications. In the sequel, we introduce two new types of MI bound that can further extend Eq. (5.1) in Theorem 5.4.3.

---

<sup>2</sup>Note that [92] uses  $I(L_i^+, L_i^-; U_i | \tilde{Z})$  but this CMI term can be strengthened to the unconditional MI by using the same development in this chapter.

### 5.4.3 Variance Based MI Bound

Inspired by the above Rademacher perspective, we first present a new bound that depends on the MI term and a notion of loss variance, defined below.

**Definition 5.4.1** ( $\gamma$ -Variance). *For any  $\gamma \in (0, 1)$ ,  $\gamma$ -variance for a learning algorithm is defined as*

$$V(\gamma) \triangleq \mathbb{E}_{W,S} \left[ \frac{1}{n} \sum_{i=1}^n (\ell(W, Z_i) - (1 + \gamma)L_S(W))^2 \right].$$

By definition,  $\gamma$ -variance also depends on the data distribution. In the zero-one loss case, it can be characterized by the following lemma.

**Lemma 5.4.3.** *Under the zero-one loss assumption, we have  $V(\gamma) = L_n - (1 - \gamma^2)\mathbb{E}_{W,S} [L_S^2(W)]$ .*

Loss variances, of any kind, have not appeared in the information-theoretic generalization bounds developed to date. Such a notion however does arise in the PAC-Bayes literature, where such an idea traces back to [185, 204]. Different from these works, here we utilize an expected empirical variance, and the distribution of  $W$  in this case is generated by the learning algorithm rather than the posterior distribution used for prediction in PAC-Bayes.

The gap between  $\mathcal{E}_\mu(\mathcal{A})$  and  $V(\gamma)$  also has a “symmetry lemma” (similar to Lemma 5.4.2) correlating to the shifted Rademacher sequence.

**Lemma 5.4.4.** *For any  $C_1 > 0$ , we have*

$$\mathcal{E}_\mu(\mathcal{A}) - C_1 V(\gamma) \leq \frac{2 + C_1 \gamma^2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i} [\tilde{\varepsilon}_i L_i^+],$$

where  $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1 \gamma^2}{C_1 \gamma^2 + 2}$  is the shifted Rademacher variable with mean  $-\frac{C_1 \gamma^2}{C_1 \gamma^2 + 2}$ .

**Theorem 5.4.4.** *Assume  $\ell(\cdot, \cdot) \in \{0, 1\}$ ,  $\gamma \in (0, 1)$ . Then, there exist  $C_1, C_2 > 0$  such that*

$$\mathcal{E}_\mu(\mathcal{A}) \leq C_1 V(\gamma) + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{n C_2}. \quad (5.5)$$

Notably, the interpolating setting is a sufficient but not necessary condition for the zero  $\gamma$ -variance, that is,  $L_n = 0$  makes  $V(\gamma) = 0$ , but  $V(\gamma) = 0$  does not indicate that  $L_n = 0$ . In addition, by Lemma 5.4.3, Eq. (5.5) can be rewritten as  $L_\mu \leq (1 + C_1)L_n - C_1(1 - \gamma^2)\mathbb{E}_{W,S} [L_S^2(W)] + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n}$  so for the fixed  $C_1$  and  $C_2$ , the bound of Eq. (5.5) is tighter than the bound of Eq. (5.1) with the gap being at least  $C_1(1 - \gamma^2)\mathbb{E}_{W,S} [L_S^2(W)]$ .

#### 5.4.4 Sharpness Based MI Bound

The nice generalization property of deep neural networks is often credited to the “flat minima” [100] of loss landscapes. Recently, [153] and [213] have proved that the generalization error can be upper-bounded by a MI based term plus a sharpness (or flatness) related term. Following the similar development in the previous section, we are able to obtain a bound that also depends on a MI term and a sharpness term, where we use a completely different analysis with [153, 213].

We first define a notion of sharpness.

**Definition 5.4.2** ( $\lambda$ -Sharpness). *For any  $\lambda \in (0, 1)$ , the “ $\lambda$ -sharpness” at position  $i$  of the training set is defined as*

$$F_i(\lambda) \triangleq \mathbb{E}_{W, Z_i} [\ell(W, Z_i) - (1 + \lambda)\mathbb{E}_{W|Z_i} \ell(W, Z_i)]^2.$$

This  $\lambda$ -sharpness can be regarded as an expected version of the “flatness” used in [233] with  $W \sim P_{W|Z_i}$  instead of some posterior distribution of  $W$ .

**Lemma 5.4.5.** *Assume  $\ell(\cdot, \cdot) \in \{0, 1\}$ , we have  $F_i(\lambda) = \mathbb{E}_{W, Z_i} [\ell(W, Z_i)] - (1 - \lambda^2)\mathbb{E}_{Z_i} [\mathbb{E}_{W|Z_i}^2 \ell(W, Z_i)]$ .*

Let  $F(\lambda) = \frac{1}{n} \sum_{i=1}^n F_i(\lambda)$ . Similar to Lemma 5.4.1, Lemma 5.4.2 and Lemma 5.4.4, we have the following symmetric argument.

**Lemma 5.4.6.** *For any  $C_1 > 0$ , we have*

$$\mathcal{E}_\mu(\mathcal{A}) - C_1 F(\lambda) = \frac{C_1 + 2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, U_i} \left[ \tilde{\varepsilon}_i L_i^+ - \frac{C_1(1 - \lambda^2)}{C_1 + 2} \hat{\varepsilon}_i h(U_i) \right],$$

where  $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1}{C_1 + 2}$  and  $\hat{\varepsilon}_i = \varepsilon_i - 1$  are the shifted Rademacher variables, and  $h(U_i) = \mathbb{E}_{\tilde{Z}_i^+ | U_i} [\mathbb{E}_{L_i^+ | \tilde{Z}_i^+, U_i}^2 L_i^+]$ .

We are then ready to present the following bound.

**Theorem 5.4.5.** *Assume  $\ell(\cdot, \cdot) \in \{0, 1\}$ ,  $\lambda \in (0, 1)$ . Then, there exist  $C_1, C_2 > 0$  such that*

$$\mathcal{E}_\mu(\mathcal{A}) \leq C_1 F(\lambda) + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n}. \quad (5.6)$$

Similar to the variance based bound, zero  $\lambda$ -sharpness is a weaker condition than the interpolating assumption. In particular, Eq. (5.6) could be tighter than Eq. (5.1) in Theorem 5.4.3 when the empirical risk is non-zero. Specifically, Eq. (5.6) can be rewritten as  $L_\mu \leq (1 + C_1)L_n - C_1(1 - \lambda^2)L_n^2 + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n}$  by Lemma 5.4.5 and Jensen's inequality. If  $C_1, C_2$  are fixed, then the sharpness based bound is always tighter than Eq. (5.1) and the gap is at least  $C_1(1 - \lambda^2)L_n^2$ .

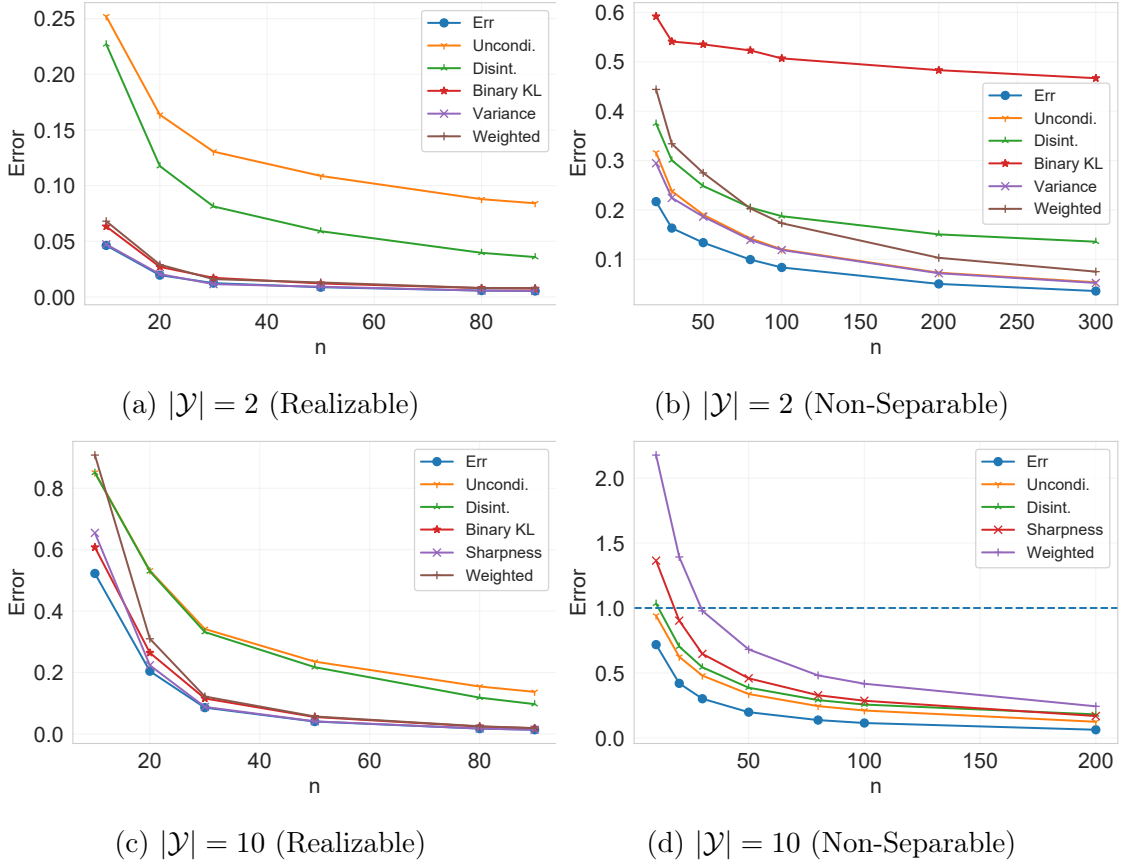


Figure 5.2: Comparison of bounds on the synthetic eCMI/dataset. (a) Binary classification with a separable  $\mu$  (i.e. the interpolating setting). Notice that the variance bound nearly coincides with  $\mathcal{E}_\mu(\mathcal{A})$ . (b) Binary classification with a non-separable  $\mu$ . (c) Ten-class classification with a separable  $\mu$ . (d) Ten-class classification with a non-separable  $\mu$ . The binary KL bound is removed in (d) since it is always  $\geq 1$ .

To conclude this section, we give a bound that combines the variance and the sharpness.

**Corollary 5.4.1.** *Assume  $\ell(\cdot, \cdot) \in \{0, 1\}$  and  $\gamma, \lambda \in (0, 1)$ , then there exist  $C_1, C_2 > 0$  such that*

$$\mathcal{E}_\mu(\mathcal{A}) \leq C_1 \min\{V(\gamma), F(\lambda)\} + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{nC_2}.$$

We remark that if  $\mathcal{A}$  satisfies any of the following: (i)  $L_n \rightarrow 0$ ; (ii)  $V(\gamma) \rightarrow 0$  for some  $\gamma \in (0, 1)$ ; (iii)  $F(\lambda) \rightarrow 0$  for some  $\lambda \in (0, 1)$ , we all have  $\mathcal{E}_\mu(\mathcal{A}) \leq \sum_{i=1}^n \frac{2I(L_i^+; U_i)}{n \ln 2}$ .

## 5.5 Numerical Results

In this section, we empirically compare some CMI and MI bounds discussed in this chapter. Our first experiment is based on a synthetic Gaussian dataset, where a simple linear classifier (with a softmax output layer) will be trained. The second experiment follows the same deep learning scenario setting with [82, 92], where we will train a 4-layer CNN on MNIST [117] and fine-tune a ResNet-50 [87] (pretrained on ImageNet [47]) on CIFAR10 [114]. In all of these experiments, we let the loss be the zero-one loss, namely  $\ell(w, z) = \mathbb{1}_{f_w(x) \neq y}$ , and we apply the empirical risk minimization (ERM) to find the hypothesis, namely  $w = \arg \min_{w \in \mathcal{W}} L_S(w)$ . Since such loss is not differentiable, to enable the gradient based optimization methods such as SGD, we hereby use the cross-entropy loss as a surrogate classification loss during training. Notice that Err is still defined with respect to the zero-one loss in our experiments.

Under these settings, we will mainly compare the disintegrated ld-CMI bound in the first inequality of Theorem 5.3.1 (*Disint.*), the unconditional MI bound in Theorem 5.3.2 (*Uncondi.*), the weighted generalization error bound in Eq. (5.2) of Theorem 5.4.3 (*Weighted*), the variance bound in Theorem 5.4.4 (*Variance*) and the sharpness bound in Theorem 5.4.5 (*Sharpness*). Besides, we will include the binary KL bound in [92] as a baseline, which is, to the best of our knowledge, the tightest fast-rate CMI bound in the literature when  $L_n$  is close (but not equal) to zero. In addition, we note that the difference between the variance bound and the sharpness bound is negligible in the current scale of the figures, so for each figure we only report one of them. The comparison between the variance bound and the sharpness bound, and more comparison of other bounds mentioned in this chapter (such as interpolating bounds and single-loss based square-root bounds) are given in Appendix C.5.



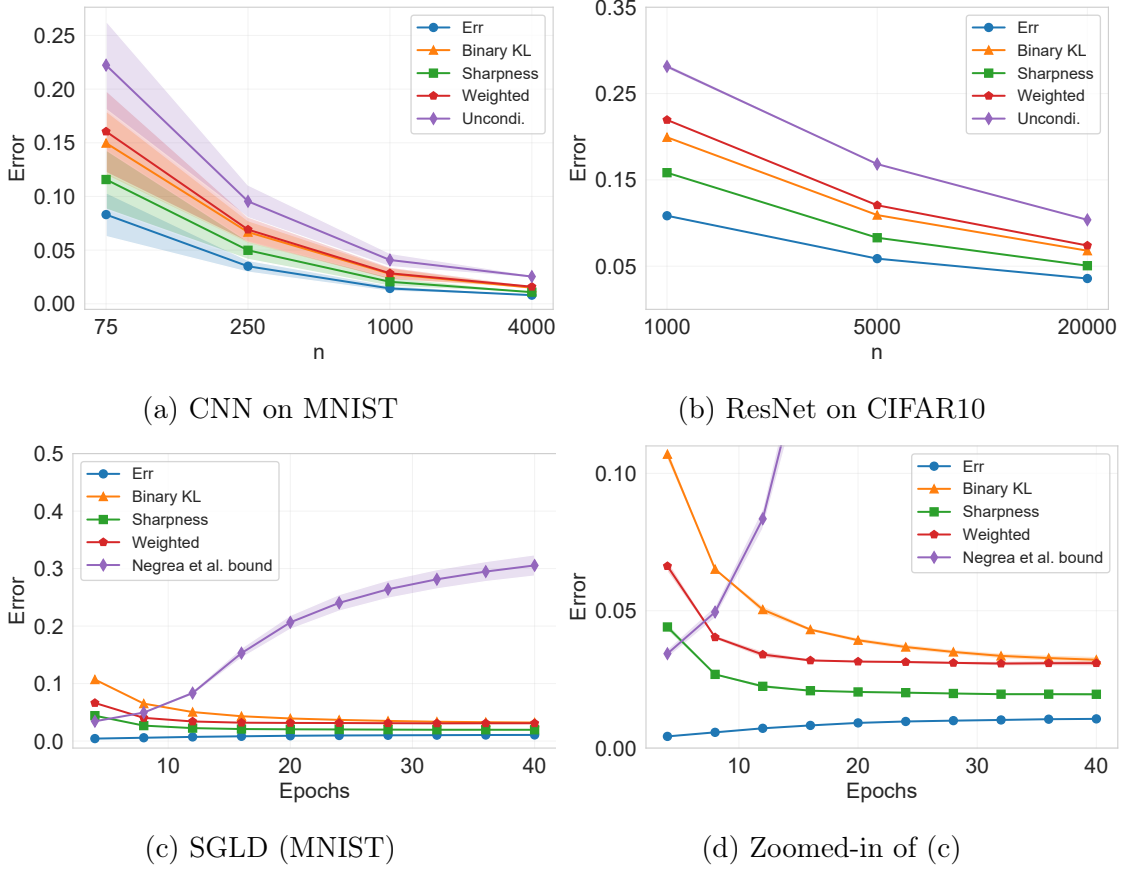


Figure 5.3: Comparison of bounds on two real datasets, MNIST (“4 vs 9”) and CIFAR10.

### 5.5.1 Linear Classifier

We will first use a simple linear classifier to carry out the Gaussian data classification task (see Appendix C.5.2 for more details of data generation and training). There are at least two major benefits of using such a synthetic dataset. On the one hand, the ground-truth distribution  $\mu$  is known so we can draw unlimited supersamples, allowing repetition of experiments so as to obtain an accurate estimate of the desired quantity (e.g., for each  $n$ , we repeat the experiment 5000 times, each with a random  $(\tilde{Z}, U)$  and report the average). On the other hand, the separability of different classes is adjustable, allowing for a control of the task difficulty. Specifically, we will consider both the zero training loss case (i.e. a separable  $\mu$ ) and the high training loss case (i.e. a non-separable  $\mu$ ).

In the binary classification tasks (i.e.  $|\mathcal{Y}| = 2$ ), the evaluations of Err and the bounds

are given in Figure 5.2a and Figure 5.2b. When  $\mu$  is separable (Figure 5.2a), the algorithm can always interpolate the training sample. In this case, the fast-rate bounds are tighter than the square-root bounds, and the variance bound (or the sharpness bound) is the tightest. Moreover, notice that here the disintegrated CMI bound is tighter than the unconditional MI bound. For a more challenging classification task (Figure 5.2b),  $L_n$  is no longer zero, the square-root bounds become tighter than the binary KL bound. Indeed, [92] shows that when the empirical risk is large, the square-root bound will be tighter than their fast-rate bounds. In contrast, our variance bound is even slightly tighter than the square-root bound of Theorem 5.3.2 in Figure 5.2b. Additionally, notice that unlike the realizable case (the one with separable  $\mu$ ), the unconditional MI bound is now tighter than the disintegrated CMI bound.

We also conduct experiments in the ten-class classification task (i.e.  $|\mathcal{Y}| = 10$ ), and the results are shown in Figure 5.2c and Figure 5.2d. In the realizable case (Figure 5.2c), the results are similar to binary classification except that the binary KL bound is tighter than all the other bounds when  $n = 10$ , which is the only case we observe where the binary KL bound outperforms Theorem 5.4.4 and Theorem 5.4.5. In addition, it is worth mentioning that Eq. (5.2) decays much faster than the square-root bounds in Figure 5.2c (and also in Figure 5.2b). For the non-separable case in Figure 5.2d, only the unconditional MI bound in Theorem 5.3.2 is non-vacuous for all the values of  $n$ . While the binary KL bound is removed in this case since it is always vacuous, our sharpness bound is competitive to the square-root bound when  $n \geq 50$ .

## 5.5.2 Neural Networks

To compare information-theoretic generalization bounds of modern deep neural networks, we follow the same experiment settings in [82, 92]. Specifically, we train a 4-layer CNN model on a binary MNIST dataset (“4 vs 9”) and also fine-tune a pretrained ResNet-50 on CIFAR10. Unlike the previous synthetic dataset case, here we can only repeatedly run experiments (with different  $\tilde{Z}$  and  $U$ ) for limited times due to the high computation complexity. Thus, we report the both average numerical values and their standard deviations. Notice that our code is primarily the same as the code provided by [92], which is originally based on the code in <https://github.com/hgrayrhar/f-CMI>. More experiment details can be found in Appendix C.5.2.

Observations in the binary MNIST experiment (Figure 5.3a) are close to the realizable binary classification case in Figure 5.2a (both near the interpolating regime). For example, the fast-rate bounds are tighter than the square-root bound. Notably, our sharpness bound

(or variance bound) significantly improve the the binary KL bound in both the MNIST experiment (Figure 5.3a) and the CIFAR10 experiment (Figure 5.3b), while Eq. (5.2) is slightly weaker than the binary KL bound. Furthermore, we also compare the bounds when the CNN model is trained by a SGLD algorithm [172], a variant of SGD, on the binary MNIST dataset. In this case, we add the weight-based MI bound of SGLD in [151, Eq. 6] as a baseline. Figure 5.3c suggests that both of our sharpness bound and Eq. (5.2) improve the binary KL bound. Notably, [82] and [92] observe that  $f$ -CMI bound and e-CMI bound are worse than the SGLD bound in [151] at the beginning of training. As shown in Figure 5.3d, although our sharpness bound is still looser than the SGLD bound in [151] before the fifth epoch, our sharpness bound significantly shrinks the gap with the SGLD bound during early training.

## 5.6 Limitations and Other Related Literature

**Limitations** More recently, the limitations of information-theoretic bounds in explaining the generalization properties of stochastic convex optimization (SCO) problems have been investigated by [78]. In their study, the authors demonstrate that almost all existing information-theoretic bounds, except for the chained MI/CMI bounds, fail to vanish in at least one of their counterexamples, despite the true generalization error vanishing. Unfortunately, neither our loss-difference MI/CMI bounds nor our single-loss MI bounds are capable of overcoming such limitations revealed in their constructed counterexample presented in [78, Theorem 17]. Note that we will tackle these limitations in Chapter 7.

In [92], the authors provide an e-CMI generalization bound for a generic convex function of the training loss and test loss. Although our analysis, using either loss-difference CMI/MI bounds or single-loss MI bounds, may not be directly applicable to general convex comparison functions between the training loss and testing loss, one potential alternative is to consider the square of the loss difference, for which similar techniques can be employed to derive generalization bounds.

Furthermore, it is important to note that all our new information-theoretic generalization bounds are derived under the assumption of independent and identically distributed (i.i.d.) training instances. Exploring the possibility of relaxing this assumption represents a promising avenue for future research.

**Other Related Work** As also mentioned in Chapter 3, information-theoretic generalization bounds have been explored for some specific algorithms. For example, the weight

based information-theoretic bounds have been successfully applied to characterize the generalization properties of SGLD [168, 29, 151, 77, 179, 210], and more recently, these bounds are also used to analyze either the vanilla SGD [153, 213] or the stochastic differential equations (SDEs) approximation of SGD [214]. To apply the weight based MI or CMI bounds for SGD and its variants, unlike the bounds in this chapter and [82, 92] that treat the learning algorithm as a black-box, these weight based bounds are usually further upper bounded by some quantities along the trajectories of the algorithms (e.g., gradient incoherence [151]). This then points to a future direction: Can the losses-based or predictions-based information-theoretic bounds be exploited the similar trajectory information of the gradient based algorithms?

It is also noteworthy that recently [76] and [175] concurrently propose a variant of the initial CMI framework [199], the “leave-one-out” (LOO) CMI setting, where their supersample is a  $n + 1$  vector instead of a  $n \times 2$  matrix. While our development in this chapter is restricted to the latter, it is curious and tempting to compare—or connect—the two. We leave this as a future work.

In addition to the supervised learning setting, information-theoretic bounds have found applicability in various other learning scenarios, showcasing their versatility. These scenarios include meta-learning [91], semi-supervised learning [86], transfer learning [225], domain adaptation [215], and so on. It is reasonable to expect that our findings can be effectively utilized in these diverse learning settings as well.

## 5.7 Conclusion

It is generally crucial to establish tighter generalization bounds that are also easy to measure. Otherwise, it becomes challenging to determine whether the insights derived from theoretical guarantees hold practical significance. In this chapter, we obtain some novel and easy-to-measure information-theoretic generalization bounds. These bounds are demonstrated to be tighter than the previous results in the same supersample setting of [199], either theoretically or empirically. In our development, we also discuss some other viewpoints of generalization in the current supersample construction including explaining generalization via the rate of reliable communication over the memoryless channel, and via correlating with the Rademacher sequence. These new insights may help to design new learning algorithms or discover novel algorithm-dependent complexity measures.

## Chapter 6

# Information-Theoretic Analysis of Unsupervised Domain Adaptation

### 6.1 Overview and Main Contributions

In this chapter, we analyze the generalization ability of hypotheses and learning algorithms for *unsupervised domain adaptation (UDA)* tasks using the information-theoretic framework. The foundation of our technique is still the Donsker-Varadhan representation of KL divergence (cf. Lemma 2.4.1). We present novel upper bounds for two notions of generalization errors. The first notion (“population-to-population (PP) generalization error”) measures the gap between the population risk in the target domain and that in the source domain *for a hypothesis*, and the second (“expected empirical-to-population (EP) generalization error”) measures the gap between the population risk in the target domain and the empirical risk in the source domain *for a learning algorithm*. We show that the PP generalization error for all hypotheses are uniformly bounded by a quantity governed by the KL divergence between the two domain distributions, which, under bounded losses, recovers the the bound in [158]. We then show that this KL term upper-bounds some other measures including Total-Variation distance [21], Wasserstein distance [190] and domain disagreement [63]. Thus, minimizing KL-divergence forces the minimization of other discrepancy measures as well. This, together with the ease of minimizing KL [158], explains the effectiveness of the KL-guided alignment approach. For expected EP generalization error, we develop several algorithm-dependent generalization bounds. These algorithm-dependent bounds further inspire the design of two new and yet simple strategies that can further boost the performance of the KL guided marginal alignment algorithms. Experiments are

performed to verify the effectiveness of these strategies.

## 6.2 Related Work

**Domain Adaptation** Many domain adaptation generalization bounds have been developed [21, 20, 46, 138, 190, 242, 63, 3], and various discrepancy measures are introduced to derive these bounds including total variation [21, 20, 46, 138], Wasserstein distance [190], domain disagreement [63] and so on. In particular, bounds based on  $\mathcal{H}\Delta\mathcal{H}$  in [20] are restricted to a binary classification setting and assume a deterministic labeling function. Furthermore, [20] also assumes the loss is the  $L_1$  distance between the predicted label and true label (which is bounded). Our bounds work for the general supervised learning problems with any labelling mechanism (e.g., stochastic labelling), and we do not require the specific choice of the loss (even unbounded). Recently, [192] proposed generalization bounds using Jensen-Shannon (JS) divergence, which bear a relation to our Corollary 6.4.2. While other algorithm-dependent bounds have been proposed for different transfer learning settings (e.g., [209]), they are not directly comparable to our own bounds. For more details about the domain adaptation theory, we refer readers to [176] for a comprehensive survey. In addition, the most common methods for domain adaptation involve aligning the marginal distributions of the representations between the source and target domains, for example, using an adversarial training mechanism [60, 190, 3] or aligning the first two moments of the representation distribution [200]. There are numerous other domain adaptation algorithms, and we refer readers to [222, 246, 211] for recent advances.

**Information-Theoretic Generalization Bounds Beyond Supervised Learning Setting** Recently, information-theoretic analysis has also been used in other learning settings including meta-learning [106, 107, 177, 35], semi-supervised learning [85, 7] and transfer learning [225, 106, 105, 139, 28]. In particular, [225, 105] consider a different problem setup with ours. Specifically, their expected generalization error is the gap between the target population risk and a weighted empirical risk combining both the source and the target empirical risks, while our “EP” error is the gap between the target population risk and the source empirical risk. That is, we focus on the role of the *unlabelled* target data in cross-domain generalization when the source empirical risk is taken as a training objective, whereas their works assume the existence of *labelled* target data and study their role in domain adaptation.

## 6.3 Refined Notation Systems

Unlike previous chapters, now we have two domains: a source domain and a target domain. Let  $\mu$  and  $\mu'$  be two distributions on  $\mathcal{Z}$ , unknown to the learner, where  $\mu$  characterizes the source domain and  $\mu'$  characterizes the target domain. We may also write  $\mu$  as  $P_Z$  or  $P_{XY}$  and  $\mu'$  as  $P_{Z'}$  or  $P_{X'Y'}$ , which defines random variables  $Z = (X, Y)$  and  $Z' = (X', Y')$ , respectively. Let  $S = \{Z_i\}_{i=1}^n \sim \mu^{\otimes n}$  be a labeled source-domain sample and  $S'_{X'} = \{X'_j\}_{j=1}^m \sim P_{X'}^{\otimes m}$  be an unlabelled target-domain sample. The objective of UDA is to design an algorithm  $\mathcal{A}$  that takes  $S$  and  $S'_{X'}$  as the input and outputs a weight  $W \in \mathcal{W}$ , giving rise to a predictor  $f_W \in \mathcal{F}$  that “works well” on the target domain. Note that the algorithm  $\mathcal{A}$  is characterized by a conditional distribution  $P_{W|S, S'_{X'}}$ .

Let  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_0^+$  be a loss function. The population risk for each  $w \in \mathcal{W}$  in the target domain is defined as

$$L_{\mu'}(w) \triangleq \mathbb{E}_{Z'}[\ell(f_w(X'), Y')]$$

and a good UDA algorithm hopes to return a weight  $w$  that minimizes this risk. Since  $\mu'$  is unknown, one often uses recourse to the empirical risk in the source domain, defined as

$$L_S(w) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(f_w(X_i), Y_i).$$

Generalization error in this setting measures how well the hypothesis returned from the algorithm generalizes from the source-domain training sample to the target-domain unknown distribution  $\mu'$ . Taking into account the stochastic nature of the algorithm  $\mathcal{A}$ , a natural notion of generalization error for UDA can be defined by

$$\text{Err} \triangleq \mathbb{E}_{W, S} [L_{\mu'}(W) - L_S(W)] = \mathbb{E}_{W, S, S'_{X'}} [L_{\mu'}(W) - L_S(W)], \quad (6.1)$$

where the expectation in the first expression is taken over the joint distribution of  $(W, S) \sim P_{W|S} \times \mu^{\otimes n}$ , and the expectation of the second expression is taken over the joint distribution of  $(W, S, S'_{X'}) \sim P_{W|S, S'_{X'}} \times \mu^{\otimes n} \times P_{X'}^{\otimes m}$ .

There is another notion of generalization error, more traditional in the domain adaptation literature, defined as the gap between the population risk in the target domain and that in the source domain:

$$\widetilde{\text{Err}}(w) \triangleq L_{\mu'}(w) - L_{\mu}(w). \quad (6.2)$$

where  $L_\mu(w) \triangleq \mathbb{E}_Z[\ell(f_w(X), Y)]$ . It is apparent that  $\widetilde{\text{Err}}(w)$  and  $\text{Err}$  are related by the following triangle inequality:

$$|L_{\mu'}(w) - L_S(w)| \leq |L_{\mu'}(w) - L_\mu(w)| + |L_\mu(w) - L_S(w)|.$$

where the second term on the right hand side is the standard generalization error in the source domain, which can be bounded by classical learning-theoretic tools, e.g., Rademacher complexity [15]. Thus, bounding  $\widetilde{\text{Err}}(w)$  helps bounding  $\text{Err}$ .

This chapter studies both notions of generalization error for UDA. Specifically, starting from Section 6.5, we will mainly use information-theoretic tools to bound  $\text{Err}$  directly, without going through  $\widetilde{\text{Err}}(w)$ . For the ease of reference, we refer to  $\widetilde{\text{Err}}(w)$  as the *population-to-population (PP) generalization error* for  $w$  and  $\text{Err}$  as the *expected empirical-to-population (EP) generalization error*.

## 6.4 Upper Bounds for PP Generalization Error

We now present some upper bounds for  $\widetilde{\text{Err}}(w)$ . The key techniques used in developing these bounds are the information-theoretic tools in the style of Lemma 2.4.1. These bounds adopt certain KL divergence to measure the discrepancy between the source and target domains. Notably, some previously established bounds are recovered under weaker conditions. Additionally, we demonstrate that under certain conditions, the KL-based bound is an upper bound of several other discrepancy measures and hence minimizing the KL divergence forces the minimization of these other measures.

We first list some common assumptions on the loss function, which we consider in this chapter.

**Assumption 6.1** (Boundedness).  $\ell(\cdot, \cdot)$  is bounded in  $[0, M]$ .

**Assumption 6.2** (Subgaussianity).  $\ell(f_w(X), Y)$  is  $R$ -subgaussian under  $\mu$  for any  $w \in \mathcal{W}$ .

**Remark 6.4.1.** Note that Assumption 6.1 implies Assumption 6.2, i.e., if  $\ell(f_w(X), Y)$  is bounded in  $[0, M]$ , then it is also  $M/2$ -subgaussian. Thus, Assumption 6.2 is weaker than Assumption 6.1.

**Assumption 6.3** (Lipschitzness).  $\ell(f_w(X), Y)$  is  $\beta$ -Lipschitz continuous in  $\mathcal{Z}$  with respect to a metric  $d$  on  $\mathcal{Z}$  for any  $w \in \mathcal{W}$ , i.e.,  $|\ell(f_w(x_1), y_1) - \ell(f_w(x_2), y_2)| \leq \beta d(z_1, z_2)$  for some metric  $d$  on  $\mathcal{Z}$ .



**Remark 6.4.2.** Note that Assumption 6.1 implies Assumption 6.3 when  $d(z_1, z_2) = \mathbb{1}_{z_1 \neq z_2}$ , i.e., if  $\ell(f_w(X), Y)$  is bounded in  $[0, M]$ , then it is also  $M$ -Lipschitz under the discrete metric.

**Assumption 6.4** (Triangle and Symmetric).  $\ell(\cdot, \cdot)$  satisfies the following:  $\ell(y_1, y_2) = \ell(y_2, y_1)$  and  $\ell(y_1, y_2) \leq \ell(y_1, y_3) + \ell(y_3, y_2)$  for any  $y_1, y_2, y_3 \in \mathcal{Y}$ .

### 6.4.1 Generalization Bounds via the Subgaussian Condition

The following generalization bound is established by combining Lemma 2.4.1 and Assumption 6.2, and its corresponding sample complexity bound is discussed in Appendix D.2.8.

**Theorem 6.4.1.** If Assumption 6.2 holds, then for any  $w \in \mathcal{W}$ ,  $\left| \widetilde{\text{Err}}(w) \right| \leq \sqrt{2R^2 \text{D}_{\text{KL}}(\mu' || \mu)}$ .

Notably this result can be turned into a generalization upper bound providing guidance to algorithm design, and at the same time it provides a lower bound of the generalization error, highlighting some fundamental difficulty of the learning task. To illustrate this, we present a corollary while noting that similar development can also be applied to other bounds presented later in this chapter.

Consider that each  $f_w$  is expressed as the composition  $g \circ h$ , where  $h$  is a function mapping  $\mathcal{X}$  to a representation space  $\mathcal{T}$  and  $g$  is a function mapping  $\mathcal{T}$  to  $\mathcal{Y}$ . For any given  $h : \mathcal{X} \rightarrow \mathcal{T}$ , denote by  $\mu_h$  the distribution on  $\mathcal{T} \times \mathcal{Y}$  obtained by pushing forward  $\mu$  via  $h$ , that is,  $\mu_h(t, y) = \int \delta(t - h(x)) d\mu(x, y)$ , where  $\delta$  is the Dirac measure on  $\mathcal{T}$ . Similarly, let  $\mu'_h$  denote the distribution on  $\mathcal{T} \times \mathcal{Y}$  obtained by pushing forward  $\mu'$  via  $h$ .

**Corollary 6.4.1.** Suppose that  $f_w = g \circ h$  and that Assumption 6.2 holds, then for any  $w \in \mathcal{W}$ ,

$$L_\mu(w) - \sqrt{2R^2 \text{D}_{\text{KL}}(\mu' || \mu)} \leq L_{\mu'}(w) \leq L_\mu(w) + \sqrt{2R^2 \text{D}_{\text{KL}}(\mu'_h || \mu_h)}.$$

In this result, the lower bound of  $L_{\mu'}(w)$  indicates a fundamental difficulty in UDA learning in that, using the same predictor mapping  $f_w$ , there is no way for the population risk in the target domain to be lower than that of the source domain less than a constant which depends only on the domain difference. On the other hand, the upper bound suggests that it is possible to squeeze the gap between the two population risks by choosing an appropriate representation map  $h$  - evidently such a map should be attempting to align  $\mu'_h$  with  $\mu_h$  or to align their respective proxies.

It is also noteworthy that under Assumption 6.1 and due to Remark 6.4.1, Theorem 6.4.1 implies

$$\left| \widetilde{\text{Err}}(w) \right| \leq \frac{M}{\sqrt{2}} \sqrt{\text{D}_{\text{KL}}(P_{X'} || P_X) + \text{D}_{\text{KL}}(P_{Y'|X'} || P_{Y|X})}. \quad (6.3)$$

Similarly applying this result in the representation space  $\mathcal{T}$ , we see that Eq. (6.3) recovers the bound in Proposition 1 of [158]. Notice that unlike [158], Theorem 6.4.1 ( or Eq. (6.3)) does not require the loss to be the cross-entropy loss.

Theorem 6.4.1 and [158] both use the KL divergence from source domain to target domain,  $\text{D}_{\text{KL}}(\mu' || \mu)$ , and in fact,  $\left| \widetilde{\text{Err}}(w) \right|$  can also be upper bounded by  $\text{D}_{\text{KL}}(\mu || \mu')$ . This can be done by invoking the subgaussianity of  $\ell(f_w(X'), Y')$  (rather than  $\ell(f_w(X), Y)$ ); for bounded loss, the subgaussianity of  $\ell(f_w(X'), Y')$  is also satisfied. Then we obtain the following corollary.

**Corollary 6.4.2.** *If Assumption 6.1 holds,  $\left| \widetilde{\text{Err}}(w) \right| \leq \frac{M}{\sqrt{2}} \sqrt{\min\{\text{D}_{\text{KL}}(\mu || \mu'), \text{D}_{\text{KL}}(\mu' || \mu)\}} \leq \frac{M}{2} \sqrt{\text{D}_{\text{KL}}(\mu || \mu') + \text{D}_{\text{KL}}(\mu' || \mu)}$ .*

**Remark 6.4.3.** *In the second inequality of Corollary 6.4.2,  $\text{D}_{\text{KL}}(\mu || \mu') + \text{D}_{\text{KL}}(\mu' || \mu)$  is known as the symmetrized KL divergence, or Jeffrey’s divergence [101], and in fact, [158] penalizes this measure between the source and target distributions in the representation space. Notice that bounds in [192] are based on the JS divergence. Since there is a sharp upper bound of the JS divergence based on Jeffrey’s divergence [43], minimizing Jeffrey’s divergence (in the representation space) will simultaneously penalize the JS divergence.*

In UDA, since  $Y'$  is completely unavailable to the algorithm  $\mathcal{A}$ , it is impossible to minimize the misalignment of conditional distributions, i.e.  $\text{D}_{\text{KL}}(P_{Y'|T'} || P_{Y|T})$  where  $T$  and  $T'$  are representations of source domain and target domain, respectively. A common method is to assign pseudo labels to target data based on a learned source classifier [126]. However, it may also cause additional issues [191]. For concreteness, suppose the trained model  $Q$  can well approximate the real mapping between  $X$  and  $Y$  on source domain (i.e.  $Q_{Y|T} = P_{Y|T}$ ), which is usually the training objective. Let  $\hat{Y}'$  be the pseudo label of  $T'$  generated by the trained model, i.e.,  $Q_{\hat{Y}'|T'} = Q_{Y|T}$ . Let  $Q_{T', \hat{Y}'} = P_{T'} Q_{\hat{Y}'|T'}$ , then the following holds,

$$\text{D}_{\text{KL}}(P_{T', Y'} || P_{T, Y}) = \mathbb{E}_{P_{T', Y'}} \log \frac{P_{T', Y'} Q_{T', \hat{Y}'}}{Q_{T', \hat{Y}'} P_{T, Y}} = \text{D}_{\text{KL}}(P_{T'} || P_T) + \text{D}_{\text{KL}}(P_{Y'|T'} || Q_{\hat{Y}'|T'}). \quad (6.4)$$

For a specific  $t'$ , if  $P(Y' = y'|T' = t') \neq 0$  and  $Q(\hat{Y}' = y'|T' = t') = 0$ , then the second term in RHS of Eq. (6.4),  $D_{\text{KL}}(P_{Y'|T'}||Q_{\hat{Y}'|T'}) \rightarrow \infty$ . In this case, even when the marginal distributions are perfectly aligned, the overall value of the upper bound is large. Thus, incorrect pseudo labels may even have negative impact on the target domain performance.

In fact, the misalignment of the conditional distributions appears to be the main difficulty of UDA [21, 3]. The next corollary suggests that this difficulty may be alleviated when the loss function satisfies the triangle property, namely, Assumption 6.4. It can be verified that this assumption is satisfied by the 0-1 loss<sup>1</sup>; this assumption has also been considered in previous works [138, 190].

**Theorem 6.4.2.** *If Assumption 6.4 holds and let  $\ell(f_{w'}(X), f_w(X))$  be  $R$ -subgaussian for any  $w, w' \in \mathcal{W}$ . Then for any  $w$ ,  $\widetilde{\text{Err}}(w) \leq \sqrt{2R^2 D_{\text{KL}}(P_{X'}||P_X)} + \lambda^*$ , where  $\lambda^* = \min_{w \in \mathcal{W}} L_{\mu'}(w) + L_{\mu}(w)$ .*

Here  $\lambda^*$  measures the possibility of whether the domain adaptation algorithm will succeed under the oracle knowledge of  $\mu$  and  $\mu'$ . In particular, if the hypothesis space is large enough, the minimizer  $w^*$  for the “joint population risk”  $L_{\mu'}(w) + L_{\mu}(w)$  may give rise to  $L_{\mu'}(w^*) = L_{\mu}(w^*) = 0$ , then we’re likely to generalize well on the target domain. Then the KL divergence  $D_{\text{KL}}(P_{X'}||P_X)$  between the two  $\mathcal{X}$ -marginals alone bounds the PP generalization error uniformly for all  $w \in \mathcal{W}$ .

This theorem motivates the strategy of penalizing  $D_{\text{KL}}(P_{T'}||P_T)$  in the representation space for UDA. The next theorem suggests that such an approach also penalizes other notions of domain discrepancy, for example, the key quantity in the PAC-Bayes type of domain adaptation generalization bounds [63], that is defined as

$$\text{dis}(P_X, P_{X'}) \triangleq |\mathbb{E}_{W, W'} [\mathbb{E}_{X'} [\ell(f_W(X'), f_{W'}(X'))]] - \mathbb{E}_{W, W'} [\mathbb{E}_X [\ell(f_W(X), f_{W'}(X))]]|. \quad (6.5)$$

**Theorem 6.4.3.** *If  $\ell(f_{w'}(X), f_w(X))$  is  $R$ -subgaussian for any  $f_w, f_{w'} \in \mathcal{F}$ , then  $\text{dis}(P_X, P_{X'}) \leq \sqrt{2R^2 D_{\text{KL}}(P_{X'}||P_X)}$ .*

Note that unlike [63], here we do not require the loss function to be the 0-1 loss.

## 6.4.2 Generalization Bounds via the Lipschitz Condition

We now present such generalization bound for UDA under the Lipschitz continuity assumption of the loss function, where  $\mathbb{W}(\cdot, \cdot)$  denotes the Wasserstein distance.

<sup>1</sup>Some losses that only satisfy a general version of Assumption 6.4 are discussed in Appendix D.2.10

**Theorem 6.4.4.** *If Assumption 6.3 holds, then  $|\widetilde{\text{Err}}(w)| \leq \beta \mathbb{W}(\mu', \mu)$ .*

Theorem 6.4.4 can be related to the KL-based bounds in the previous section when the Wasserstein distance is defined with respect to the discrete metric  $d$ . In this case and under bounded loss function, which is also Lipschitz continuous, Theorem 6.4.4 follows. On the other hand, Wasserstein distance is also equivalent to the total variation in this case, while the latter is connected to the KL divergence via Pinsker’s inequality [170, Theorem 6.5] and the Bretagnolle-Huber inequality [27, Lemma 2.1]. Thus, we arrive at the following result.

**Corollary 6.4.3.** *If Assumption 6.1 holds and let  $d$  be the discrete metric, then*

$$|\widetilde{\text{Err}}(w)| \leq M \text{TV}(\mu', \mu) \leq M \sqrt{\min \left\{ \frac{1}{2} \text{D}_{\text{KL}}(\mu' || \mu), 1 - e^{-\text{D}_{\text{KL}}(\mu' || \mu)} \right\}}.$$

Note that results here are inspired by the work of [180]. Corollary 6.4.3 provides a tighter bound than the one in Eq. (6.3), as can be directly verified.

Parallel to Theorem 6.4.2, if the loss function satisfies the triangle property, we may establish the bound below, which recovers a similar result in [190, Theorem 1.] but without restricting the task to be binary classification or requiring the loss to be the  $L_1$  distance.

**Theorem 6.4.5.** *If Assumption 6.4 holds and  $\ell(f_w(X), f_{w'}(X))$  is  $\beta$ -Lipschitz in  $\mathcal{X}$  for any  $w, w' \in \mathcal{W}$ , then for any  $w \in \mathcal{W}$ ,  $\widetilde{\text{Err}}(w) \leq \beta \mathbb{W}(P_{X'}, P_X) + \lambda^*$ , where  $\lambda^* = \min_{w \in \mathcal{W}} L_{\mu'}(w) + L_{\mu}(w)$ .*

These results justify the strategy of minimizing domain discrepancy in the representation space. Since the KL-based bounds upper-bound those based on other measures of domain differences, penalizing the KL divergence will also penalize those other measures. This is practically advantageous since it is usually easier and more stable to minimize the KL divergence [158].

## 6.5 Upper Bounds for EP Generalization Error and Applications

There are two limitations in the bounds on the PP generalization error developed so far and in the traditional analysis of UDA. First, such bounds are independent of  $w$  and

hence algorithm-independent. Second, although these bounds may inspire strategies to exploit the unlabelled target sample, e.g., aligning the source and target distributions in the representation space, they only provide very limited knowledge on the role that the unlabelled target sample plays. Inspired by the works of [151] and [179], we derive upper bounds for the EP generalization error that take better advantage of the dependence of the algorithm's output on the unlabelled target data. Applications of these bounds in designing the learning algorithms are also presented.

### 6.5.1 EP Generalization Bounds

**Theorem 6.5.1.** *Assume  $\ell(f_W(X'), Y')$  is  $R$ -subgaussian under  $P_{W, Z'_j | X'_j = x'_j}$  for any  $x'_j \in \mathcal{X}$ , then*

$$|\text{Err}| \leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{2R^2 I^{X'_j}_j(W; Z_i)} + \sqrt{2R^2 \text{D}_{\text{KL}}(\mu || \mu')}.$$

**Remark 6.5.1.** *It is worth noting that the unlabelled target data contributes to the first term of the bound. Increasing the amount of source and target data will result in a reduction of the first term in the bound. Specifically, moving the expectation inside the square root function by Jensen's inequality and since  $Z_i \perp\!\!\!\perp X'_j$ , the equations  $I(W; Z_i | X'_j) = I(W; Z_i | X'_j) + I(Z_i; X'_j) = I(W; Z_i) + I(X'_j; Z_i | W)$  hold by the chain rule. The term  $I(W; Z_i)$  will vanish as  $n \rightarrow \infty$  and the term  $I(X'_j; Z_i | W)$  will also vanish as  $n, m \rightarrow \infty$ .*

The theorem can be turned into a version that is more practically relevant, in which the KL term is replaced with their representation-space counter-part (following a similar argument used for deriving Corollary 6.4.1). In addition, note that although larger sample sizes allow better estimation of that KL term, utilizing pseudo-labels for estimation may have a negative impact (as discussed in Section 6.4), which can be amplified by the larger sample size.

**Corollary 6.5.1.** *Let Assumption 6.1 hold. Then*

$$|\text{Err}| \leq \frac{M}{\sqrt{2nm}} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{\min \{I^{X'_j}_j(W; Z_i), L^{X'_j}_j(W; Z_i)\}} + \frac{M}{\sqrt{2}} \sqrt{\min \{\text{D}_{\text{KL}}(\mu || \mu'), \text{D}_{\text{KL}}(\mu' || \mu)\}}.$$

**Theorem 6.5.2.** *Assume  $\ell$  is Lipschitz for both  $w \in \mathcal{W}$  and  $z \in \mathcal{Z}$ , i.e.,  $|\ell(f_w(x), y) - \ell(f_w(x'), y')| \leq \beta d_1(z, z')$  for all  $z, z' \in \mathcal{Z}$  and  $|\ell(f_w(x), y) - \ell(f_{w'}(x), y)| \leq \beta' d_2(w, w')$  for all  $w, w' \in \mathcal{W}$ , then*

$$|\text{Err}| \leq \frac{\beta'}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j, Z_i} \mathbb{W}(P_{W|Z_i, X'_j}, P_{W|X'_j}) + \beta \mathbb{W}(\mu, \mu').$$

This bound is tighter than the bound in Theorem 6.5.1, as can be indicated by the following corollary.

**Corollary 6.5.2.** *Let Assumption 6.1 hold. Then*

$$\begin{aligned} |\widetilde{\text{Err}}| &\leq \frac{M}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j, Z_i} \left[ \text{TV}(P_{W|Z_i, X'_j}, P_{W|X'_j}) \right] + M \text{TV}(\mu, \mu') \\ &\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j, Z_i} \sqrt{\frac{M^2}{2} \text{D}_{\text{KL}}(P_{W|Z_i, X'_j} \| P_{W|X'_j})} + \sqrt{\frac{M^2}{2} \text{D}_{\text{KL}}(\mu \| \mu')}. \end{aligned}$$

Notice that to recover Theorem 6.5.1 from Corollary 6.5.2 (under Assumption 6.1), we can use Jensen's inequality to move the expectation over  $Z_i$  to inside the square root function.

## 6.5.2 Gradient Penalty as an Universal Regularizer

The algorithm-dependent bound in Theorem 6.5.1 tells us that one can reduce the EP error by limiting the disintegrated mutual information  $I^{X'_j}(W; Z_i)$ . In the stochastic gradient based optimization algorithms, this term can be controlled by penalizing the gradient norm. To see this, we now consider a “noisy” iterative algorithm for updating  $W$ , e.g., SGLD. At each time step  $t$ , let the labelled mini-batch from the source domain be  $Z_{B_t}$ , let the unlabelled mini-batch from the target domain be  $X'_{B_t}$ , and let  $g(W_{t-1}, Z_{B_t}, X'_{B_t})$  be the gradient at time  $t$ . Thus, the updating rule of  $W$  is  $W_t = W_{t-1} - \eta_t g(W_{t-1}, Z_{B_t}, X'_{B_t}) + N_t$  where  $\eta_t$  is the learning rate and  $N_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$  is an isotropic Gaussian noise. Inspired by [168], we have the following bound.

**Theorem 6.5.3.** *Let the total iteration number be  $T$  and let  $G_t = g(W_{t-1}, Z_{B_t}, X'_{B_t})$ , then*

$$|\text{Err}| \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\eta_t^2}{\sigma_t^2} \mathbb{E}_{S'_{X'}, W_{t-1}, S} \left[ \|G_t - \mathbb{E}_{Z_{B_t}} [G_t]\|^2 \right]} + \sqrt{2R^2 \text{D}_{\text{KL}}(\mu \| \mu')}.$$

**Remark 6.5.2.** *Considering a noisy iterative algorithm here is merely for simplifying analysis. In fact, it is also possible to analyze the original iterative gradient optimization method without noise injected. For example, one can follow the same development in [153, 213] to analyze vanilla SGD. In that case, there will be some residual terms in the bound.*

Theorem 6.5.3 hints that to reduce the generalization error, one can simply restrict the gradient norm at each step (so that  $\|G_t - \mathbb{E}_{Z_{B_t}}[G_t]\|^2$  is reduced). This strategy will also restrict the distance between the final output  $W_T$  and the initialization  $W_0$ , effectively shrinking the hypothesis space accessible by the algorithm. We also note that the importance of gradient penalty has been theoretically justified in the supervised learning setting [151, 77, 197, 179, 153, 213, 214].

Indeed, adding gradient penalty can be applied to any existing UDA algorithm and it is simple but effective in practice. Later on we will show that even when the algorithm  $\mathcal{A}$  does not access to any target data, in which case  $I(W; Z_i | X'_j)$  reduces to  $I(W; Z_i)$  and  $g(W_{t-1}, Z_{B_t}, X'_{B_t})$  becomes  $g(W_{t-1}, Z_{B_t})$ , minimizing the empirical loss of source domain sample while penalizing gradient norm will still improve the performance. Notice that gradient penalty has been used in standard supervised learning as a regularization technique [61, 99]. It is also used in Wasserstein distance based adversarial adaptation [69, 190], and their motivation is to stabilize the training to avoid gradient vanishing problem. Here we suggest, with strong theoretical justification, that gradient penalty is a universal technique for improving the generalization performance in UDA for any gradient-based learning method.

Notably the bound in Theorem 6.5.3 only depends on the size  $n$  of labelled source sample and does not explicitly depend on  $m$ , the size of unlabelled target sample. With a more careful design, if we consider the mutual information as the expected KL divergence of a posterior and a prior, based on  $I^{X'_j}(W; Z_i)$  in Theorem 6.5.1, it is possible to create a target-data-dependent prior and derive a tighter bound based on some quantity similar to “gradient incoherence” in [151].

### 6.5.3 Controlling Label Information for KL Guided Marginal Alignment

Consider instances in the representation space,  $Z = (T, Y)$  and  $Z' = (T', Y)$ . Theorem 6.5.1 also encourages us to align the distributions of two domains in the representation space, as argued earlier. Then the KL guided marginal alignment algorithm proposed in [158] can be invoked here. One may notice that Theorem 6.5.1 uses  $D_{\text{KL}}(\mu || \mu')$  while [158] uses  $D_{\text{KL}}(\mu' || \mu)$ . As already discussed in Section 6.4, this inconsistency can be ignored when the loss is bounded (see Corollary 6.5.1).

Most domain adaptation algorithms aim to align the marginal distributions of two domains in the representation space. However, without accessing to  $Y'$ , it remains unknown if an UDA algorithm will work well since we cannot guarantee that discrepancy

between conditional distribution  $P_{Y|T}$  and  $P_{Y'|T'}$  won't become too large when we align the marginals. In [158], the authors show that  $D_{\text{KL}}(P_{Y'|T'}||P_{Y|T})$  can be upper-bounded by  $D_{\text{KL}}(P_{Y'|X'}||P_{Y|X})$ , if  $I(X;Y) = I(T;Y)$ . The authors then argue that penalizing the KL divergence of the marginals is safe.

We now argue that in practice the condition  $I(X;Y) = I(T;Y)$  can be difficult to satisfy if the cross-entropy loss is used to define the source-domain empirical risk.

By data processing inequality on  $Y - X - T$ , we know that  $I(X;Y) \geq I(T;Y) = H(Y) - H(Y|T)$ . Thus, to let  $I(T;Y)$  reach its maximum, one must minimize  $H(Y|T)$ . On the other hand, let  $Q_{Y|T,W}$  be the predictive distribution of labels in the source domain generated by the classifier. The expected cross-entropy loss for each  $Z_i$  in the representation space is then

$$\mathbb{E}_{W,Z_i} [\ell(f_W(T_i), Y_i)] = \mathbb{E}_{Z_i} [\mathbb{E}_{W|Z_i} [-\log Q_{Y_i|T_i,W}]],$$

which also decomposes as [2, 83]

$$\mathbb{E}_{W,Z_i} [\ell(f_W(T_i), Y_i)] = H(Y_i|T_i) + \mathbb{E}_{T_i,W} [D_{\text{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})] - I(W;Y_i|T_i). \quad (6.6)$$

Then minimizing the expected cross-entropy loss may not adequately reduce  $H(Y_i|T_i)$  but rather cause  $I(W;Y_i|T_i)$  to significantly increase, particularly when the model capacity is large. This may have two negative effects. First, the condition  $I(X;Y) = I(T;Y)$  is significantly violated, and  $D_{\text{KL}}(P_{Y'|T'}||P_{Y|T})$  is no longer upper bounded by  $D_{\text{KL}}(P_{Y'|X'}||P_{Y|X})$ . Hence, aligning the two marginals alone may not be adequate. Second, large  $I(W;Y_i|T_i)$  indicates  $W$  just simply memorizes the label  $Y_i$ , resulting a form of overfitting and hurting the generalization performance.

The key take-away from the above analysis is that when aligning the marginals in UDA, controlling the source label information in the weights can be important to achieve good cross-domain generalization. A similar message can also be deduced from Theorem 6.5.1, when it is viewed in the representation space and noting  $I^{T'_j}(W;Z_i) = I^{T'_j}(W;T_i) + I^{T'_j}(W;Y_i|T_i)$ .

To control label information, [83] proposed an approach called LIMIT. However, this method is rather complicated and arguably hard to train in domain adaptation (see Appendix D.3.8). We now derive a simple alternative strategy for this purpose.

Notice that  $I^{T'_j}(W;Y_i|T_i) \leq \inf_Q \mathbb{E}_{T_i} [D_{\text{KL}}(P_{W|Y_i,T_i,T'_j=t'_j}||Q_{W|T_i,T'_j=t'_j})]$ , which is a simple extension of variational representation of mutual information [170, Corollary 3.1.]. Here  $Q$  could be any distribution. By assuming  $P = \mathcal{N}(W, \sigma^2 \mathbf{I}_d | Y_i, T_i, T'_j = t'_j)$  and taking  $Q = \mathcal{N}(\widetilde{W}, \tilde{\sigma}^2 \mathbf{I}_d | T_i, T'_j = t'_j)$ , we have

$$I^{T'_j}(W;Y_i|T_i) \leq \inf_Q \mathbb{E}_{T_i} [D_{\text{KL}}(P_{W|Y_i,T_i,T'_j=t'_j}||Q_{\widetilde{W}|T_i,T'_j=t'_j})] \propto \|W - \widetilde{W}\|^2.$$



Table 6.1: RotatedMNIST and Digits. Results of baselines are reported from [158].

| Method | RotatedMNIST ( $0^\circ$ as source domain) |                                |                                |                                |                                |             | Digits                              |                                     |                                     |             |
|--------|--------------------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|-------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------|
|        | $15^\circ$                                 | $30^\circ$                     | $45^\circ$                     | $60^\circ$                     | $75^\circ$                     | Ave         | $\mathbf{M} \rightarrow \mathbf{U}$ | $\mathbf{U} \rightarrow \mathbf{M}$ | $\mathbf{S} \rightarrow \mathbf{M}$ | Ave         |
| ERM    | 97.5 $\pm$ 0.2                             | 84.1 $\pm$ 0.8                 | 53.9 $\pm$ 0.7                 | 34.2 $\pm$ 0.4                 | 22.3 $\pm$ 0.5                 | 58.4        | 73.1 $\pm$ 4.2                      | 54.8 $\pm$ 6.2                      | 65.9 $\pm$ 1.4                      | 64.6        |
| DANN   | 97.3 $\pm$ 0.4                             | 90.6 $\pm$ 1.1                 | 68.7 $\pm$ 4.2                 | 30.8 $\pm$ 0.6                 | 19.0 $\pm$ 0.6                 | 61.3        | 90.7 $\pm$ 0.4                      | 91.2 $\pm$ 0.8                      | 71.1 $\pm$ 0.5                      | 84.3        |
| MMD    | 97.5 $\pm$ 0.1                             | 95.3 $\pm$ 0.4                 | 73.6 $\pm$ 2.1                 | 44.2 $\pm$ 1.8                 | 32.1 $\pm$ 2.1                 | 68.6        | 91.8 $\pm$ 0.3                      | 94.4 $\pm$ 0.5                      | 82.8 $\pm$ 0.3                      | 89.7        |
| CORAL  | 97.1 $\pm$ 0.3                             | 82.3 $\pm$ 0.3                 | 56.0 $\pm$ 2.4                 | 30.8 $\pm$ 0.2                 | 27.1 $\pm$ 1.7                 | 58.7        | 88.0 $\pm$ 1.9                      | 83.3 $\pm$ 0.1                      | 69.3 $\pm$ 0.6                      | 80.2        |
| WD     | 96.7 $\pm$ 0.3                             | 93.1 $\pm$ 1.2                 | 64.1 $\pm$ 3.3                 | 41.4 $\pm$ 7.6                 | 27.6 $\pm$ 2.0                 | 64.6        | 88.2 $\pm$ 0.6                      | 60.2 $\pm$ 1.8                      | 68.4 $\pm$ 2.5                      | 72.3        |
| KL     | 97.8 $\pm$ 0.1                             | 97.1 $\pm$ 0.2                 | 93.4 $\pm$ 0.8                 | 75.5 $\pm$ 2.4                 | 68.1 $\pm$ 1.8                 | 86.4        | 98.2 $\pm$ 0.2                      | 97.3 $\pm$ 0.5                      | 92.5 $\pm$ 0.9                      | 96.0        |
| ERM-GP | 97.5 $\pm$ 0.1                             | 86.2 $\pm$ 0.5                 | 62.0 $\pm$ 1.9                 | 34.8 $\pm$ 2.1                 | 26.1 $\pm$ 1.2                 | 61.2        | 91.3 $\pm$ 1.6                      | 72.7 $\pm$ 4.2                      | 68.4 $\pm$ 0.2                      | 77.5        |
| KL-GP  | 98.2 $\pm$ 0.2                             | 96.9 $\pm$ 0.1                 | 95.0 $\pm$ 0.6                 | <b>88.0<math>\pm</math>8.1</b> | <b>78.1<math>\pm</math>2.5</b> | <b>91.2</b> | 98.8 $\pm$ 0.1                      | <b>97.8<math>\pm</math>0.1</b>      | <b>93.8<math>\pm</math>1.1</b>      | <b>96.8</b> |
| KL-CL  | <b>98.4<math>\pm</math>0.2</b>             | <b>97.3<math>\pm</math>0.2</b> | <b>95.6<math>\pm</math>0.1</b> | 83.0 $\pm$ 8.2                 | 73.6 $\pm$ 4.0                 | 89.6        | <b>98.9<math>\pm</math>0.1</b>      | 97.7 $\pm$ 0.1                      | 93.0 $\pm$ 0.3                      | 96.5        |

Thus, we may create an auxiliary classifier  $f_{\tilde{w}}$  that is not allowed to access to the real source label  $Y$ . In each iteration, we use the pseudo labels of target data (and source data) assigned by  $f_w$  to train  $f_{\tilde{w}}$  and adding  $\|W - \tilde{W}\|^2$  as a regularizer in the training of  $W$ . The algorithm is given in the Appendix. Remarkably the regularizer here resembles “Projection Norm” designed in [237] for out-of-distribution generalization.

## 6.6 Experimental Results

We perform experiments to verify the proposed techniques inspired by our theory.

**Datasets** We select two popular small datasets, RotatedMNIST and Digits, to compare the different methods. RotatedMNIST is built based on the MNIST dataset [117] and consists of six domains, each containing 11,666 images. These six domains are rotated MNIST images with rotation angle  $0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ$  and  $75^\circ$ , respectively. We will take the original MNIST dataset ( $0^\circ$ ) as the source domain and take other five domains as target domains. Hence, there are five domain adaptation tasks on RotatedMNIST. Digits consists of three sub-datasets, namely MNIST, USPS [97] and SVHN [152], and the corresponding domain adaptation tasks are MNIST $\rightarrow$ USPS ( $\mathbf{M}\rightarrow\mathbf{U}$ ), USPS $\rightarrow$ MNIST ( $\mathbf{U}\rightarrow\mathbf{M}$ ), SVHN $\rightarrow$ MNIST ( $\mathbf{S}\rightarrow\mathbf{M}$ ).

**Compared Methods** Baseline methods are some popular marginal alignment UDA methods including **DANN** [60], **MMD** [120], **CORAL** [200], **WD** [190] and **KL** [158]. We also choose **ERM** as another baseline, in which only the source-domain sample is

accessible during training. To verify the strategies inspired by our theory, we first add the gradient penalty to the ERM algorithm (**ERM-GP**), and we then combine gradient penalty (GP) and controlling label information (CL) with the recent proposed KL guided marginal alignment method, which are denoted by **KL-GP** and **KL-CL**, respectively.

**Implementation Details** Most of our implementation is based on the *DomainBed* suite [70]. Other settings exactly follow [158] and the results of baseline methods are taken from [158]. Specifically, each algorithm is run three times and we show the average performance with the error bar. Every dataset has a validation set, and the model selection scheme is based on the best performance achieved on the validation set of target domain during training (oracle). The hyper-parameter searching process is also built upon the implementation in the *DomainBed* suite. Other details and additional experiments can be found in Appendix D.

**Results** From Table 6.1, we first notice that gradient penalty allows **ERM** to perform more comparably to other marginal alignment methods. For example, on RotatedMNIST, **ERM-GP** outperforms **CORAL** and performs nearly the same with **DANN**. On Digits, **ERM-GP** outperforms **WD**. When GP and CL combined with KL guided algorithm, we can see that the performance can be further boosted. This justifies the discussion in Section 6.5.2 and Section 6.5.3.

## 6.7 Conclusion

Despite that the numerous learning techniques have been developed for domain adaptation, significant room exists for more in-depth theoretical understanding and more principled design of learning algorithms. This chapter presents the information-theoretic analysis for unsupervised domain adaptation, where we query two notions of the generalization errors in this context and present novel learning bounds. Some of these bounds recover the previous KL-based bounds under different conditions and confirm the insights in the learning algorithms that align the source and target distributions in the representation space. Our other bounds are algorithm-dependent, better exploiting the unlabelled target data, which have inspired novel and yet simple schemes for the design of learning algorithms. We demonstrate the effectiveness of these schemes on standard benchmark datasets.

# Chapter 7

## Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds

### 7.1 Overview and Main Contributions

Information-theoretic generalization bounds have been the subject of extensive discussions for their two main deficiencies. The first deficiency concerns the unbounded nature of the original input-output mutual information (IOMI) bounds [232, 17, 132, 131]. To address this issue, several techniques have been developed, including the chaining method [11], the individual technique [29] or random subset technique [151], and the Gaussian noise perturbation technique [153]. Notably, the CMI bound [199], as introduced in Chapter 5, stands out as it has a finite upper bound for any learning scenario due to its supersample construction. These techniques are now often applied jointly for analyzing generalization [77, 179, 247, 82, 92, 217]. The second deficiency concerns the sub-optimal convergence rate of information-theoretic bounds. Specifically, when the information-theoretic quantities, whether IOMI or CMI, are bounded by constants, the bounds exhibit a decaying rate in the order of  $\mathcal{O}(1/\sqrt{n})$ , where  $n$  is the sample size. In contrast, generalization errors in practice may decay at a faster rate, e.g.,  $\mathcal{O}(1/n)$ . To address this limitation, several works, inspired by some PAC-Bayesian literature, have proposed fast-rate information-theoretic bounds [90, 92, 217], demonstrating a good characterization in some instances of non-

convex settings such as deep learning. More recently, [226, 227, 249] have presented IOMI bounds for the Gaussian mean estimation problem, that achieve optimal convergence rates, contrasting previous bounds [29, 247].

The issue of slow convergence in information-theoretic bounds has recently been amplified by the observation that these bounds may not even vanish. [78] highlight this limitation in the context of stochastic convex optimization (SCO) problems [187]. Specifically, [78] shows that all existing information-theoretic bounds, including the CMI bound [199], the Gaussian noise perturbed IOMI bound [153], the individual IOMI bound [29], the evaluated CMI (e-CMI) bound [199] and the functional CMI ( $f$ -CMI) bound [82], fail to vanish in at least one of the counterexamples they constructed. These failures stem from the dimension-dependent nature of current information-theoretic quantities [131], appearing an intrinsic barrier for overcoming these limitations. However, if we dissect the process by which these bounds are derived, opportunities do exist. Specifically, recall that all these information-theoretic bounds are built upon the Donsker-Varadhan (DV) variational representation of the KL divergence [170, Theorem 3.5] (see Lemma 2.4.1). Using this representation, a generalization upper bound is derived in terms of information-theoretic quantity and a cumulant-generating function (CGF), as illustrated below.

$$\text{Generalization Error} \leq \inf_{\text{Para.} > 0} \frac{\text{IOMI or CMI} + \text{CGF}}{\text{Para.}}.$$

Particularly note that the CGF depends on certain choice of the auxiliary function in the DV formula. Then, except for the IOMI or CMI itself, the tightness of the generalization bound hinges on two key factors: the selection of the DV auxiliary function and the approach used to bound the CGF, the latter of which often requires additional assumptions specific to the chosen DV auxiliary function. The most common choices for the DV auxiliary function involve making assumptions such as subgaussian loss or bounded loss. For instance, in the case of IOMI bounds, the DV auxiliary function is typically chosen as the single loss or average loss, and the subgaussian assumption (or boundedness assumption) is utilized. In CMI bounds, the DV auxiliary function is defined as the difference in loss between a training sample and a “ghost” sample, and the boundedness property is employed. We now note that exploiting bounded loss is the fundamental reason behind the failures in SCO problems. Although [78] does not explicitly rely on boundedness, they do make use of the product of the Lipschitz constant and the diameter of the hypothesis domain, which essentially serves as an upper bound for the loss. The product does not vanish with  $n$ , thereby neutralizing the potential to include another decaying factor in the bound. Arguably, at least for these SCO problems, the selection of the DV auxiliary function may not be optimal, for example, the resulting  $\text{CGF} = \mathcal{O}(1/n)$ , giving vacuous generalization

guarantees. This analysis inspires us to explore alternative DV auxiliary function and to adopt different assumptions for bounding the CGF. For instance, we may devise appropriate stability assumptions and create opportunities to upper bound the CGF using a term in  $\mathcal{O}(\beta^2/n)$ , where  $\beta$  is the stability parameter that decays as  $n$ . This is promising in light of the capability of the stability-based framework in explaining the generalization of SCO problems [24, 188, 80, 16].

In this chapter, we combine information-theoretic analysis with stability notions to develop IOMI bounds and CMI bounds that improve upon previous information-theoretic bounds for stable learning algorithms, achieving faster convergence rates. The main contributions of this chapter are summarized below.

- We introduce our new notions of algorithmic stability, referred to as sample-conditioned hypothesis (SCH) stability. We also present a novel construction of a sample-dependent hypothesis matrix, where each column is a neighborhood pair of hypotheses obtained from two training samples that differ in only one element, inspired by the supersample setting of CMI [199].
- We present new IOMI bounds, which explicitly include the SCH stability parameters and are shown superior to previous bounds for stable learning algorithms.
- We show that the sample-dependent hypothesis matrix, similar to the supersample matrix, enjoys a symmetry property. Exploiting this symmetry, we establish novel CMI bounds. Specifically, we present hypotheses-conditioned CMI bounds that are analogous to the previous supersample-conditioned CMI bounds. Additionally, we derive sample-conditioned CMI bounds exploiting other assumptions. Notably, these bounds introduce novel CMI quantities and include SCH stability parameters. In particular, the new CMI quantities remain boundedness as the original CMI. Consequently, these CMI bounds vanish no slower than their stability parameters. In addition, we also obtain a second-moment generalization bound that matches the tightest known bound in the literature [54] under the same condition.
- We apply our new bounds to a convex-Lipschitz-bounded (CLB) example in which previous information-theoretic bounds fail to explain generalization [78], and show that the new IOMI and CMI bounds vanish, benefiting from SCH stability. Additionally, we discuss another CLB example where the uniform stability parameter is non-vanishing or has a slow convergence rate, but our information-theoretic generalization bounds remain tight up to a constant.

- We extend our analysis to derive information-theoretic generalization bounds based on Bernstein condition, which leverages a connection between stability and Bernstein condition. We also show that stability can be incorporated into generalization bounds to obtain stronger results using alternative information-theoretic quantities such as the loss-difference-based CMI, e-CMI and  $f$ -CMI. Furthermore, we illustrate the expressiveness of our new CMI notions under the distribution-free setting.

## 7.2 Other Related Literature

The connection between information-theoretic bounds, including some PAC-Bayes bounds, and algorithmic stability has been explored in previous literature [173, 133, 178, 121, 199, 82, 12, 175]. These works primarily focus on either regarding the information-theoretic quantities as notions of distributional stability [173, 232, 121, 199, 12] and/or converting information-theoretic quantities to some other algorithmic stability notions [199, 82, 175]. The later often relies on the addition of Gaussian noise to the hypotheses (or assuming that the prior and posterior distributions are Gaussian in PAC-Bayes). In [133], the authors also combine the DV formula with stability assumptions, where they derive some PAC-Bayes bounds. However, comparing these bounds with others is challenging in general due to the presence of their hyperparameter stability. We note that even the advanced fast-rate PAC-Bayes bound in [68] still fails to vanish in the certain example of SCO [78, Theorem 9].

## 7.3 Additional Preliminaries

**Uniform Stability** Let  $R \in \mathcal{R}$  be a source of randomness (a random variable independent of  $S$  over an appropriate space  $\mathcal{R}$ ), from which a learning algorithm  $\mathcal{A} : \mathcal{Z}^n \times \mathcal{R} \rightarrow \mathcal{W}$  takes the training sample  $S$  and  $R$  as input, and outputs a hypothesis  $W = \mathcal{A}(S, R) \in \mathcal{W}$ .

We now give two notions of uniform stability [24, 52], where  $s \simeq s^i$  denotes two training sets  $s$  and  $s^i$  that differ only at the  $i$ th element. We say a learning algorithm  $\mathcal{A}$  is  $\beta_1$ -weakly uniformly stable if

$$\sup_{s \simeq s^i, z} \mathbb{E}_R |\ell(\mathcal{A}(s, R), z) - \ell(\mathcal{A}(s^i, R), z)| \leq \beta_1,$$

and  $\beta_2$ -strongly uniformly stable if

$$\sup_{s \simeq s^i, z} \sup_r |\ell(\mathcal{A}(s, r), z) - \ell(\mathcal{A}(s^i, r), z)| \leq \beta_2.$$

**Remark 7.3.1.** *Notably the weak uniform stability above is the standard in the literature. It is evident that if  $\mathcal{A}$  is  $\beta$ -strongly uniformly stable, it must also be  $\beta$ -weakly uniformly stable. It is also worth noting that for deterministic algorithms (e.g., GD or fixed permutation SGD), the two notions are identical. We note that in this chapter, when speaking of uniform stability, we refer to the strong notion.*

Let  $S' = \{Z'_i\}_{i=1}^n \sim \mu^n$  be an independent copy of  $S$ , and let  $S^{\setminus i} = S \setminus \{Z_i\}$  so  $S^i = S^{\setminus i} \cup \{Z'_i\}$ .

The following well-known result (e.g., [24, Lemma 7], [186, Thm. 13.2]) is frequently used in this chapter.

**Lemma 7.3.1.** *For any algorithm  $\mathcal{A}$ , we have  $L_\mu = \mathbb{E}_{S,S',R} \left[ \frac{1}{n} \sum_{i=1}^n \ell(\mathcal{A}(S^i, R), Z_i) \right]$ , and*

$$\mathcal{E}_\mu(\mathcal{A}) = \mathbb{E}_{S,S'} \left[ \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{\mathcal{A}(S^i, R) | S, Z'_i} \ell(\mathcal{A}(S^i, R), Z_i) - \mathbb{E}_{\mathcal{A}(S, R) | S} \ell(\mathcal{A}(S, R), Z_i) \right] \right]. \quad (7.1)$$

**Supersample and Sample-Conditioned Hypothesis Stability** Following [199] and Chapter 5, let  $\tilde{Z} \in \mathcal{Z}^{n \times 2}$  be a supersample matrix with  $n$  rows and 2 columns, each entry drawn independently from  $\mu$ . We index the columns of  $\tilde{Z}$  by 0, 1 and denote the  $i$ th row of  $\tilde{Z}$  by  $\tilde{Z}_i$ , with entries  $(\tilde{Z}_{i,0}, \tilde{Z}_{i,1})$ . We often use the superscripts  $+$  and  $-$  to respectively replace the subscripts 0 and 1, i.e., writing  $\tilde{Z}_i^+$  for  $\tilde{Z}_{i,0}$  and  $\tilde{Z}_i^-$  for  $\tilde{Z}_{i,1}$ . Correspondingly,  $\tilde{Z}_{[n]}^+$  and  $\tilde{Z}_{[n]}^-$  denote the first and second columns, respectively. Additionally, we will use  $\tilde{Z}_{[n] \sim i}^+$  to denote  $\tilde{Z}_{[n]}^+$  in which the  $i$ th element  $\tilde{Z}_i^+$  is replaced with the corresponding element  $\tilde{Z}_i^-$  in the second column.

Let  $\tilde{W}^+ = \mathcal{A}(\tilde{Z}_{[n]}^+, R)$ , and  $\tilde{W}_i^- = \mathcal{A}(\tilde{Z}_{[n] \sim i}^+, R)$  for each  $i \in [n]$ . That is,  $\tilde{W}^+$  and each  $\tilde{W}_i^-$  are obtained by two “neighboring” training samples, namely those differ only on one instance (the  $i$ th instance). We then construct matrix  $\tilde{W} \in \mathcal{W}^{n \times 2}$ , where the  $i$ th row  $\tilde{W}_i = (\tilde{W}_i^+, \tilde{W}_i^-) = (\tilde{W}_{i,0}, \tilde{W}_{i,1})$  and  $\tilde{W}_i^+ = \tilde{W}^+$  for all  $i \in [n]$ , as shown in Table 7.1. Unlike the supersample matrix, elements in  $\tilde{W}$  are identically distributed but not independent. In this case, Eq. (7.1) can be rewritten as

$$\begin{array}{|c|c|} \hline \tilde{Z}_1^+ & \tilde{Z}_1^- \\ \hline \tilde{Z}_2^+ & \tilde{Z}_2^- \\ \hline \vdots & \vdots \\ \hline \tilde{Z}_n^+ & \tilde{Z}_n^- \\ \hline \end{array} \xRightarrow{\mathcal{A}} \begin{array}{|c|c|} \hline \tilde{W}^+ & \tilde{W}_1^- \\ \hline \tilde{W}^+ & \tilde{W}_2^- \\ \hline \vdots & \vdots \\ \hline \tilde{W}^+ & \tilde{W}_n^- \\ \hline \end{array}$$

Table 7.1: Supersample (Left) and the induced hypotheses (Right).

$$\mathcal{E}_\mu(\mathcal{A}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i^+, \tilde{W}_i} \left[ \ell(\tilde{W}_i^-, \tilde{Z}_i^+) - \ell(\tilde{W}_i^+, \tilde{Z}_i^+) \right]. \quad (7.2)$$

The summand in Eq (7.2) exhibits an interesting “symmetry”: for any  $i$ ,

$$\mathbb{E}_{\tilde{Z}_i^+, \tilde{W}_i} \left[ \ell(\tilde{W}_i^-, \tilde{Z}_i^+) - \ell(\tilde{W}_i^+, \tilde{Z}_i^+) \right] = \mathbb{E}_{\tilde{Z}_i^-, \tilde{W}_i} \left[ \ell(\tilde{W}_i^+, \tilde{Z}_i^-) - \ell(\tilde{W}_i^-, \tilde{Z}_i^-) \right]. \quad (7.3)$$

That is, the  $+/-$  superscripts in the summand of Eq (7.2) can be flipped for any  $i$ . We may then use  $n$  binary ( $\{0, 1\}$ -valued) variables  $(U_1, U_2, \dots, U_n) := U$  to govern whether we choose to flip the signs for each of the  $n$  terms, where  $U_i = 0$  indicates “no flipping”. Let  $\bar{U}_i \triangleq 1 - U_i$ , we have

$$\begin{aligned} \mathcal{E}_\mu(\mathcal{A}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i, \tilde{W}_i, U_i} \left[ \ell(\tilde{W}_{i, \bar{U}_i}, \tilde{Z}_{i, U_i}) - \ell(\tilde{W}_{i, U_i}, \tilde{Z}_{i, U_i}) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i, \tilde{W}_i, U_i} \left[ (-1)^{U_i} \left( \ell(\tilde{W}_i^-, \tilde{Z}_{i, U_i}) - \ell(\tilde{W}_i^+, \tilde{Z}_{i, U_i}) \right) \right] \end{aligned} \quad (7.4)$$

$$= \mathbb{E}_{\tilde{W}, E, U} \left[ \frac{1}{n} \sum_{i=1}^n (-1)^{U_i} \left( \ell(\tilde{W}_i^-, \hat{Z}_i) - \ell(\tilde{W}_i^+, \hat{Z}_i) \right) \right], \quad (7.5)$$

where in Eq. (7.4) we have chosen  $U$  to be an i.i.d.  $\text{Bern}(\frac{1}{2})$  sequence, and in Eq. (7.5) we have renamed  $\tilde{Z}_{i, U_i}$  as  $\hat{Z}_i$  and denoted  $E \triangleq (\hat{Z}_1, \hat{Z}_2, \dots, \hat{Z}_n)$ . Note that  $E$ , induced by  $U$  from  $\tilde{Z}$ , contains  $n$  instances, each serving to evaluate the loss difference between a pair of hypotheses  $(\tilde{W}_i^-, \tilde{W}_i^+)$ . The polarities of these evaluations are governed by  $U$ .

We now define some new notions of stability.

**Definition 7.3.1** (Sample-Conditioned Hypothesis (SCH) Stability). *Let  $S \sim \mu^n$ ,  $W$  and  $W^i$  generated via  $W = \mathcal{A}(S, R)$  and  $W^i = \mathcal{A}(S^i, R)$ . Let  $\mathcal{Z}_{w, w^i}$  denote the support of the conditional distribution  $P_{Z_i|w, w^i}$ , and  $\mathcal{W}_s$  denote the support of the conditional distribution  $P_{W|s}$ . We introduce four types of SCH stability, referred to types A, B, C, D. Specifically, a learning algorithm  $\mathcal{A}$  is*

$$a) \gamma_1\text{-SCH-A stable if } \forall i \in [n], \quad \sup_{w \in \cup_{s \in \mathcal{Z}^n} \mathcal{W}_s} \sup_{z \in \mathcal{Z}} \left| \ell(w, z) - \mathbb{E}_{W^i|w} [\ell(W^i, z)] \right| \leq \gamma_1, \quad (7.6)$$

$$b) \gamma_2\text{-SCH-B stable if } \forall i \in [n], \quad \mathbb{E}_{S, R, Z'} \left[ \left( \ell(W, Z') - \mathbb{E}_{W^i|W} [\ell(W^i, Z')] \right)^2 \right] \leq \gamma_2^2, \quad (7.7)$$

$$c) \gamma_3\text{-SCH-C stable if } \forall i \in [n], \quad \mathbb{E}_{W, W^i} \left[ \sup_{z_i \in \mathcal{Z}_{W, W^i}} \left| \ell(W, z_i) - \ell(W^i, z_i) \right| \right] \leq \gamma_3, \quad (7.8)$$

$$d) \gamma_4\text{-SCH-D stable if } \forall i \in [n], \quad \mathbb{E}_{S, Z'_i, R} \left[ \left( \ell(W, Z_i) - \ell(W^i, Z_i) \right)^2 \right] \leq \gamma_4^2, \quad (7.9)$$



where  $Z'$  in Eq. (7.7) is an independent instance drawn from  $\mu$ .

**Remark 7.3.2.** By definition, we can see  $\gamma_2 \leq \gamma_1$ , and it is expected that  $\gamma_4 \leq \gamma_3$ . Note that all of them are smaller than  $\beta_2$ . In addition, it is expected that  $\gamma_2$ ,  $\gamma_3$  and  $\gamma_4$  are smaller than  $\beta_1$ , although the relationship between  $\gamma_1$  and  $\beta_1$  is uncertain. Moreover, it is also expected that  $\gamma_4$  is larger than  $\gamma_2$  due to the independence of  $Z'$  in Eq. (7.7). We emphasize that supreme  $\sup_w$  in Eq. (7.6) is taken over the sample-dependent hypothesis space  $\cup_{s \in \mathcal{Z}^n} \mathcal{W}_s$ , not the whole hypothesis space  $\mathcal{W}$ . Notably, the notion of “hypothesis set stability” in [57] is closely related to our SCH stability. In fact,  $\gamma_1$ -SCH-A stable implies  $\gamma_1$ -hypothesis set stable. Further elaboration on the definition of SCH stability can be found in Appendix E.2.

## 7.4 IOMI Bounds for Stable Algorithms

We are now in a position to give the IOMI bounds for stable learning algorithms.

**Theorem 7.4.1.** *If the learning algorithm  $\mathcal{A}$  is  $\gamma_1$ -SCH-A stable, then*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\sqrt{2}\gamma_1}{n} \sum_{i=1}^n \sqrt{I(\widetilde{W}^+; \widetilde{Z}_i^+)} \leq \frac{\sqrt{2}\gamma_1}{n} \sum_{i=1}^n \sqrt{I(\widetilde{W}^+; \widetilde{Z}_i^+ | \widetilde{W}_i^-)}.$$

We note that  $I(\widetilde{W}^+; \widetilde{Z}_i^+) = I(\widetilde{W}_i^-; \widetilde{Z}_i^-) = I(W; Z_i)$ , the only difference between the first bound in Theorem 7.4.1 and the previous individual IOMI bound in [29] is that the subgaussian variance proxy is replaced by the stability parameter  $\gamma_1$  in our bound. In fact, while we use  $\widetilde{W}$  to better understand the appearance of  $\gamma_1$  in the IOMI bounds, the first bound in Theorem 7.4.1 itself does not necessarily rely on the supersample and the construction of  $\widetilde{W}$ .

In addition, if  $\mathcal{A}$  is a deterministic algorithm, the term  $I(\widetilde{W}^+; \widetilde{Z}_i^+ | \widetilde{W}_i^-)$  in the second bound is tighter than the mutual information stability or erasure information studied in [173, 82], that is,  $I(\widetilde{W}^+; \widetilde{Z}_i^+ | \widetilde{W}_i^-) \leq I(\widetilde{W}^+; \widetilde{Z}_i^+ | \widetilde{Z}_{[n] \setminus i}^+)$  (see Remark E.3.1 in the Appendix for an explanation).

*Proof Sketch of Theorem 7.4.1.* Motivated by Lemma 7.3.1 and Eq. (7.2), the main innovation in this proof is to let the auxiliary function in DV be the “relative loss” instead of the single loss, namely we let  $g(\tilde{w}_i^+, \tilde{z}_i^+) = \mathbb{E}_{\widetilde{W}_i^- | \tilde{w}^+} [\ell(\widetilde{W}_i^-, \tilde{z}_i^+)] - \ell(\tilde{w}^+, \tilde{z}_i^+)$  and let the auxiliary function  $f = t \cdot g$  for  $t > 0$  in Lemma 2.4.1. This enables us to utilize Eq. (7.6)

to bound the CGF. The remaining steps are routine. The complete proof can be found in Appendix E.3.1.  $\square$

The following corollary, immediately following from  $\gamma_1 \leq \beta_2$ , suggests a clear improvement over the IOMI bound for the uniformly stable deterministic algorithm with vanishing  $\beta_2$ .

**Corollary 7.4.1.** *If  $\mathcal{A}$  is  $\beta_2$ -uniform stable, then  $|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\sqrt{2}\beta_2}{n} \sum_{i=1}^n \sqrt{I(\widetilde{W}^+; \widetilde{Z}_i^+)}$ .*

A key message conveyed in this chapter, as shown in the proof of Theorem 7.4.1, is the importance of carefully selecting the DV auxiliary function and appropriate assumptions for various learning scenarios. Corollary 7.4.1 also suggests a potential enhancement for uniform stability in the certain deterministic setting, provided that  $I(W; Z_i)$  also vanishes appropriately with  $n$ .

Similar to [151, 179, 82, 92], we also give an  $R$ -conditioned IOMI bound below.

**Theorem 7.4.2.** *If  $\mathcal{A}$  is  $\beta_2$ -uniform stable, we have  $|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\sqrt{2}\beta_2}{n} \sum_{i=1}^n \mathbb{E}_R \sqrt{I^R(\widetilde{W}^+; \widetilde{Z}_i^+)}$ .*

This disintegrated IOMI bound is not directly comparable to the bounds in Theorem 7.4.1, but it is equivalent to Corollary 7.4.1 for deterministic algorithms. With additional assumptions, we may remove the square-root in Theorem 7.4.1, accelerating the decay of IOMI as shown in Theorem 7.4.3.

**Theorem 7.4.3.** *Under the same conditions in Theorem 7.4.1, if  $\mathcal{A}$  is further  $\gamma_2$ -SCH-B stable, then*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\gamma_1}{n} \sum_{i=1}^n I(\widetilde{W}^+; \widetilde{Z}_i^+) + 0.72 \frac{\gamma_2^2}{\gamma_1}.$$

Notice that  $\gamma_2^2/\gamma_1 \leq \gamma_1^2/\gamma_1 = \gamma_1$ . If  $\gamma_2^2/\gamma_1^2$  decays faster than  $\frac{1}{n} \sum_{i=1}^n \sqrt{I(\widetilde{W}^+; \widetilde{Z}_i^+)}$ , then Theorem 7.4.3 is qualitatively strictly stronger than Theorem 7.4.1.

## 7.5 CMI Bounds for Stable Algorithms

**Hypotheses-Conditioned CMI Bounds** In this section, we give a handful of CMI bounds based on a new information-theoretic quantity.

**Theorem 7.5.1.** *Suppose that there exists  $\Delta_1 : \mathcal{W}^2 \rightarrow \mathbb{R}$  such that for every  $\tilde{w}_i = (\tilde{w}_i^+, \tilde{w}_i^-)$ ,  $\sup_{z_i \in \mathcal{Z}_{\tilde{w}_i^+, \tilde{w}_i^-}} |\ell(\tilde{w}_i^+, z_i) - \ell(\tilde{w}_i^-, z_i)| \leq \Delta_1(\tilde{w}_i)$ . Then,*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\sqrt{2}}{n} \sum_{i=1}^n \min \left\{ \mathbb{E}_{\tilde{W}_i} \left[ \Delta_1(\tilde{W}_i) \sqrt{I^{\tilde{W}_i}(\hat{Z}_i; U_i)} \right], \sqrt{\mathbb{E}_{\tilde{W}_i} [\Delta_1(\tilde{W}_i)^2] I(\hat{Z}_i; U_i | \tilde{W}_i)} \right\}.$$

Furthermore, if  $\mathcal{A}$  is  $\gamma_3$ -SCH-C stable, then such a  $\Delta_1$  exists and

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\sqrt{2}\gamma_3}{n} \sum_{i=1}^n \sqrt{\sup_{\tilde{w}_i} I^{\tilde{w}_i}(\hat{Z}_i; U_i)}.$$

Compared to the standard supersample-conditioned CMI bound [199] in terms of  $I(W; U | \tilde{Z})$ , which assesses how well one can infer the “training-set membership” from the output hypothesis, our new CMI quantity, namely  $I(\hat{Z}_i; U_i | \tilde{W}_i)$ , measures our ability to decide if an instance  $\hat{Z}_i$  contributes to the training of  $\tilde{W}_i^+$  or to the training of  $\tilde{W}_i^-$  when we know it contributes to only one of them. When  $\tilde{W}_i^+$  and  $\tilde{W}_i^-$  are similar, this decision (i.e., determining  $U_i$ ) is difficult, giving rise to small  $I(\hat{Z}_i; U_i | \tilde{W}_i)$ . Additionally, for uniformly stable algorithms,  $\sup_{w_i} \Delta_1(\tilde{w}_i) \leq \beta_2$  and it is also expected that  $\mathbb{E}_{\tilde{W}_i} [\Delta_1(\tilde{W}_i)^2] \leq \beta_1^2$ . Moreover, if we simply replace  $\gamma_3$  by an upper bound of  $\ell$ , the second bound of Theorem 7.5.1 becomes a *uniform convergence* bound if  $I^{\tilde{w}}(\hat{Z}_i; U_i)$  vanishes with  $n$ .

*Proof Sketch of Theorem 7.5.1.* Again we find motivation in Lemma 7.3.1. Additionally, the symmetry exhibited in Eq. (7.5), analogous to the symmetry between  $\tilde{Z}_i^+$  and  $\tilde{Z}_i^-$  used in deriving the standard CMI bounds, allows a similar development. Specifically, letting  $g(\tilde{w}_i, \hat{z}_i, u_i) = (-1)^{u_i} (\ell(\tilde{w}_i^-, \hat{z}_i) - \ell(\tilde{w}_i^+, \hat{z}_i))$  and  $f = t \cdot g$  for  $t > 0$  in Lemma 2.4.1 enables the bounding of the CGF via invoking the assumptions of the theorem. The complete proof is given in Appendix E.4.1.  $\square$

Notably, our new CMI quantity in the bound preserves the boundedness of the original CMI in [199], that is,  $I(\hat{Z}_i; U_i | \tilde{W}) \leq H(U_i) = \log 2$ . Furthermore, parallel to supersample-conditioned CMI being smaller than IOMI [77], a similar result for hypotheses-conditioned CMI is given below.

**Theorem 7.5.2.** *For any  $\mathcal{A}$  and  $\mu$ , we have  $I(\hat{Z}_i; U_i | \tilde{W}_i) \leq I(W; Z_i)$ .*

Similar to Theorem 7.4.3, we present a CMI bound without square-root, which could be much stronger in certain regimes.

**Theorem 7.5.3.** Let  $\Delta_1$  be defined in the same way as in Theorem 7.5.1, and we let  $\Lambda(\tilde{w}_i) = \mathbb{E}_{\hat{Z}_i|\tilde{w}_i} \left[ \left( \ell(\tilde{w}_i^-, \hat{Z}_i) - \ell(\tilde{w}_i^+, \hat{Z}_i) \right)^2 \right] / \Delta_1(\tilde{w}_i)^2$ , then

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{W}_i} \left[ \Delta_1(\tilde{W}_i) \left( I^{\tilde{W}_i}(\hat{Z}_i; U_i) + 0.72\Lambda(\tilde{W}_i) \right) \right]. \quad (7.10)$$

If  $\mathcal{A}$  is further  $\beta_2$ -uniform stable and  $\gamma_4$ -SCH-D stable, then

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\beta_2}{n} \sum_{i=1}^n I(\hat{Z}_i; U_i | \tilde{W}_i) + 0.72 \frac{\gamma_4^2}{\beta_2}. \quad (7.11)$$

Note that it is valid to set  $\gamma_3 = \mathbb{E}_{\tilde{W}_i} [\Delta_1(\tilde{W}_i)]$ , which can be viewed as a generalization bound in its own right. Additionally, it can be verified that  $\Lambda(\tilde{w}_i) \leq 1$  for any  $\tilde{w}_i$ . As  $I^{\tilde{W}_i}(\hat{Z}_i; U_i) \leq \log 2 \approx 0.69$ , Eq. (7.10) can be further upper bounded by  $1.41\gamma_3$ . This ensures that Eq. (7.10) will decay no slower than  $\mathcal{O}(\gamma_3)$ .

We also present a second moment bound.

**Theorem 7.5.4.** Assume that  $\mathcal{A}$  is  $\beta_2$ -uniform stable and symmetric with respect to  $S$ , i.e. it does not depend on the order of the elements in  $S$ . Let  $\ell(\cdot, \cdot) \in [0, 1]$ , then

$$\mathbb{E}_{W,S} [(L_\mu(W) - L_S(W))^2] \leq 4\beta_2^2 \left( \frac{1.5I(E; U|\tilde{W}) + 0.82}{n} + 1 \right) + \frac{1}{n}.$$

Since  $I(E; U|\tilde{W}) \leq \mathcal{O}(n)$ , the bound can be further upper bounded by  $\mathcal{O}(\beta_2^2 + 1/n)$ , which matches the previous tight bound for the second moment generalization error in [54, Theorem 1.2]. Notice that [54, Theorem 1.2] only holds for the deterministic setting, while our bound also holds for randomized algorithms, and we also give a stronger result in Appendix E.4.4 based on our SCH stability notions.

**Supersample-Conditioned CMI Bounds** It is also possible to give  $\tilde{Z}$ -conditioned CMI bounds that explicitly contain the stability parameters. Let  $W_i = \tilde{W}_{i,U_i}$  and  $\bar{W}_i = \tilde{W}_{i,\bar{U}_i}$ , we have the following results.

**Theorem 7.5.5.** Suppose there exists  $\Delta_2 : \mathcal{Z} \rightarrow \mathbb{R}$  such that  $\sup_{w, w^i \in \mathcal{W}_{z_i}^2} |\ell(w, z_i) - \ell(w^i, z_i)| \leq \Delta_2(z_i)$  for every  $z_i$ , where  $\mathcal{W}_{z_i}^2$  is the support of the conditional distribution  $P_{W,W^i|z_i}$ . Then,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\sqrt{2}}{n} \sum_{i=1}^n \min \left\{ \mathbb{E}_{\tilde{Z}_i^+} \left[ \Delta_2(\tilde{Z}_i^+) \sqrt{I^{\tilde{Z}_i^+}(W_i, \bar{W}_i; U_i)} \right], \sqrt{\mathbb{E}_{\tilde{Z}_i^+} [\Delta_2(\tilde{Z}_i^+)^2] I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)} \right\}.$$

The proof is deferred to Appendix E.4.5. The interpretation of the CMI quantity  $I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)$  appears identical to our earlier CMI quantity  $I(\tilde{Z}_i; U_i | \bar{W}_i)$ . A closer look in fact reveals that the two quantities are mathematically equal. To see this, first note  $I(\tilde{Z}_i; U_i | \bar{W}_i) = H(U_i) - H(U_i | \tilde{Z}_i, \bar{W}_i)$  and  $I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+) = H(U_i) - H(U_i | \tilde{Z}_i^+, W_i, \bar{W}_i)$ . Let random variable  $A_1 = (\tilde{Z}_i, \tilde{W}_i^+, \tilde{W}_i^-)$  and let  $A_2 = (\tilde{Z}_i^+, W_i, \bar{W}_i)$ , these two random variables are identically distribution (since  $P_{\tilde{W}_i^+} = P_{\tilde{W}_i^-}$ ,  $P_{\tilde{Z}_i^+} = P_{\tilde{Z}_i^-}$  and  $U_i \sim \text{Bernoulli}(\frac{1}{2})$ ) and also we have  $P_{A_1|U_i} = P_{A_2|U_i}$  so  $P_{A_1, U_i} = P_{A_2, U_i}$ , which gives us  $H(U_i | A_1) = H(U_i | A_2)$ . This indicates that  $I(\tilde{Z}_i; U_i | \tilde{W}_i) = I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)$ .

In Theorem 7.5.5, a data-dependent hypothesis space  $\mathcal{W}_{z_i}^2$  is defined. A similar concept has been utilized in the hypothesis set cross-validation (CV) stability studied in [57]. Furthermore, [57] derives some bounds based on either their transductive Rademacher complexity or their hypothesis set CV stability. They show that these two notions dominate in different learning scenarios. Given the close relationship between the supersample construction and the Rademacher complexity [199, 217], and the inspiration behind our  $\tilde{W}$  construction, our framework is likely to have a fundamental connection to [57]. Additionally, obtaining the similar results of  $\tilde{Z}$ -conditioned CMI as in Theorem 7.5.3-7.5.4 is warranted, which may require some stability notions analogous to the average CV-stability in [57].

## 7.6 Convex–Lipschitz–Bounded (CLB) Problems

We now discuss two examples of the convex-Lipschitz-bounded (CLB) problem, a subclass of SCO problems.

The first example is previously given in [78, Thm. 17], in which nearly all previous information-theoretic generalization bounds are non-vanishing. We will demonstrate that our CMI bounds are non-vacuous in this example.

**Example 7.1.** Let  $d \in \mathbb{N}$  and  $\mathcal{Z} = \{e(i) : i \in [d]\}$  where  $e(i)$  is a one-hot vector with 1 at the  $i$ -th coordinate. Let  $\mu = \text{Unif}(\mathcal{Z})$ . Given a sample  $S = \{Z_i\}_{i=1}^n$  drawn i.i.d. from  $\mu$ , we choose the 1-Lipschitz convex loss function  $\ell(w, z) = -\langle w, z \rangle$  and use GD to select a hypothesis  $w$  from  $\mathcal{W} = \{w \in \mathbb{R}^d : \|w\| \leq 1\}$ . Let the number of GD iterations be  $T = n^2$  and let the learning rate be  $\eta = \frac{1}{n\sqrt{n}}$ .

Let  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n z_i$  be the sample mean. In this deterministic setting, it's easy to see

that

$$w_t = \begin{cases} \eta t \hat{\mu} & \text{if } \eta t \|\hat{\mu}\| \leq 1, \\ \eta t \hat{\mu} / \|\eta t \hat{\mu}\| & \text{otherwise.} \end{cases}$$

Let  $\hat{\mu}^i$  be the sample mean of  $s^i$  and let  $w_t^i$  be its corresponding hypothesis at time  $t$ . Notice that Euclidean projection does not increase the distance between projected points, namely non-expansive [80, Lemma 4.6]. Hence, whether  $w_t = \eta t \hat{\mu}$  or its truncated version  $\eta t \hat{\mu} / \|\eta t \hat{\mu}\|$  limited within the unit ball, we have  $\|w_t - w_t^i\| \leq \|\eta t \hat{\mu} - \eta t \hat{\mu}^i\| \leq \mathcal{O}(\eta t/n)$ . Recall that the loss function is 1-Lipschitz, we have  $|\mathcal{E}_\mu(\mathcal{A})| \leq \beta_2 \leq \mathcal{O}(\eta t/n)$ . In this example,  $\eta T/n = 1/\sqrt{n}$  so  $\beta_2 \in \mathcal{O}(1/\sqrt{n})$ . One can also directly obtain this rate from [80, 16].

Now following the same setting in [78], if we let  $d = 2n^2$ , we can find that  $I(W_T; Z_i) \in \Omega(1)$  (see [78, Thm. 17] or Appendix E.6). Thus, IOMI itself could not explain the generalization of GD in this problem. Furthermore, all our CMI quantities including those in Section 7.7 also have the order of  $\Omega(1)$ , that is, they fail to vanish as  $n \rightarrow \infty$  (see Appendix E.6 for more elaboration).

Therefore, the stability parameter  $\beta_2$  should not be replaced by some constant (e.g., the upper bound of the loss function) in the IOMI or CMI bound. In fact, for our CMI bounds, due to their boundedness property, we have the following corollary.

**Corollary 7.6.1.** *If  $\mathcal{A}$  is  $\beta_2$ -uniform stable, we have  $\frac{\beta_2}{n} \sum_{i=1}^n \sqrt{I(\hat{Z}_i; U_i | \widetilde{W}_i)} \leq \mathcal{O}(\beta_2)$ .*

Corollary 7.6.1 provides a solution to the non-vanishing limitation of the previous information-theoretic generalization bounds in Example 7.1 (and also the counterexample in [78, Thm. 4]), as it can explain generalization as long as the stability-based bound is sufficient. Thus, the shortfalls of information-theoretic generalization bounds in analyzing deterministic algorithms for CLB problems are tempered by the stability-based framework.

We now show another CLB Example from [78, Thm. 3] where  $\mathcal{A}$  is not uniformly stable, and we will see information-theoretic generalization bounds in this paper are tight up to a constant. This example is also studied in [161, Sec. 5].

**Example 7.2.** *Let  $\mathcal{W} \in \mathbb{R}^d$  be a ball with radius  $R_0$ , and let the input space be  $\mathcal{Z} = \{z_0/R_0, -z_0/R_0\}$  where  $z_0 \in \mathcal{W}$  such that  $\|z_0\| = R_0$ . Let  $\mu = \text{Unif}(\mathcal{Z})$ . Consider a convex and  $L$ -Lipschitz loss function  $\ell(w, z) = -L\langle w, z \rangle$ . In addition,  $\mathcal{A}$  is any empirical risk minimization (ERM) algorithm.*

In this example,  $\beta_2 = 2LR_0$  is a constant. [78] has shown that  $|\mathcal{E}_\mu(\mathcal{A})| \geq \frac{LR_0}{\sqrt{2n}}$  and  $I(W; S) \leq 1$  (see [78, Thm. 3] or Appendix E.6 for an explanation). This gives us

$\frac{2LR_0}{n} \sum_{i=1} \sqrt{I(W; Z_i)} \leq 2LR_0 \sqrt{\frac{I(W; S)}{n}} \leq \frac{2LR_0}{\sqrt{n}}$ . Thus, the distribution-dependent property of IOMI can improve the stability-based bound in this case. In addition, we know that  $I(\widehat{Z}_i; U_i | \widetilde{W}_i) \leq I(W; Z_i)$  from Theorem 7.5.2, our new CMI bounds are also tight (up to a constant) in this example.

Notably, we can construct an additional example building upon Example 7.2, where  $\mathcal{A}$  is uniformly stable but the uniform stability itself results in a slow convergence rate for generalization error. Specifically, let  $R_0 = \frac{1}{d}$  and let  $d = \sqrt{n}$ , then  $\beta_2 \in \mathcal{O}(1/\sqrt{n})$  while  $|\mathcal{E}_\mu(\mathcal{A})| \geq \frac{L}{\sqrt{2n}}$ . Note that the information-theoretic bounds in this paper can still provide a tight rate, namely  $\mathcal{O}(1/n)$ .

These examples demonstrate that our bounds can improve both the stability-based bound and information-theoretic generalization bounds in some learning scenarios.

Additional applications are discussed in Appendix E.7.

## 7.7 Extensions

**Connection with Bernstein Condition** The Bernstein condition is commonly used to derive fast-rate generalization bound for both PAC-Bayes bounds [241, 33, 143, 68] and stability-based bounds [112], then it is natural to explore the relationship between our fast-rate bounds and the Bernstein condition, formally defined below.

**Definition 7.7.1** (Bernstein Condition). *Assume that  $w^* = \arg \min_{w \in \mathcal{W}} L_\mu(w)$  is a risk minimizer. We say that the Bernstein assumption is satisfied with some  $B > 0$  and  $\kappa \in [1, +\infty)$  if for any  $w \in \mathcal{W}$ ,*

$$\mathbb{E}_Z [(\ell(w, Z) - \ell(w^*, Z))^2] \leq B (L_\mu(w) - L_\mu(w^*))^{\frac{1}{\kappa}}.$$

This condition can be easily satisfied in many common situations [6, 112]. In the following proposition, we can see that the Bernstein condition implies the  $\gamma_2$ -SCH-B stable algorithm.

**Proposition 7.7.1.** *If the Bernstein condition is satisfied with some  $B$  and  $\kappa$ , then  $\mathcal{A}$  is  $\gamma_2$ -SCH-B stable, where  $\gamma_2^2 = 4B\mathbb{E}_W [(L_\mu(W) - L_\mu(w^*))^{\frac{1}{\kappa}}]$ .*

Therefore, invoking the Bernstein condition, we can obtain the fast-rate bound as presented in Theorem 7.4.3, where we need to assume the loss is bounded, as shown below as a by-product.

**Corollary 7.7.1.** *If the Bernstein condition is satisfied with  $\kappa = 1$  and  $\ell \in [0, C]$ , then*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{C}{n} \sum_{i=1}^n I(W; Z_i) + \frac{2.88B}{C} (L_\mu - L_\mu(w^*)).$$

Recently, [226, 227] use the unexpected excess risk as the DV auxiliary function and invoke the  $(\eta, c)$ -central condition to establish some optimal-rate bounds for specific learning problems, e.g., Gaussian mean estimation. This again underscores the significance of selecting appropriate DV auxiliary functions and corresponding assumptions tailored to different learning problems. It is worth mentioning that the Bernstein condition also implies their  $(\eta, c)$ -central condition. Therefore, unifying these conditions can be considered as a potential avenue for future research.

**Loss Difference, Evaluated, and Functional CMI for Stable Algorithms** Similar to [217], we derive some tighter bounds based on the loss difference.

**Theorem 7.7.1.** *Let  $\Delta L_i = \ell(W_i, \tilde{Z}_i^+) - \ell(\bar{W}_i, \tilde{Z}_i^+)$ . If  $\mathcal{A}$  is  $\beta_2$ -uniform stable, then*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\sqrt{2}\beta_2}{n} \sum_{i=1}^n \min \left\{ \sqrt{I(\Delta L_i; U_i)}, \mathbb{E}_{\tilde{Z}_i^+} \sqrt{I^{\tilde{Z}_i^+}(\Delta L_i; U_i)} \right\} \leq \frac{\sqrt{2}\beta_2}{n} \sum_{i=1}^n \sqrt{I(\Delta L_i; U_i | \tilde{Z}_i^+)}.$$

We note that while it is feasible to replace  $\beta_2$  in the (disintegrated) CMI bounds above with certain sample-conditioned hypothesis stability, it is not possible to apply the same substitution for the unconditional MI bound in Theorem 7.7.1.

Furthermore, notice that  $\Delta L_i - (L_i, \bar{L}_i) - (F_i, \bar{F}_i) - (W_i, \bar{W}_i)$  forms a Markov chain given  $\tilde{Z}_i^+$ , wherein  $(L_i, \bar{L}_i)$  are the loss pair evaluated at  $\tilde{Z}_i^+$  using  $(W_i, \bar{W}_i)$ , and  $(F_i, \bar{F}_i)$  are label predictions of  $\tilde{Z}_i^+$  using  $(W_i, \bar{W}_i)$ . By the data-processing inequality, one can obtain e-CMI bound [199, 92],  $f$ -CMI bound [82] and recover  $I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)$  based bound from Theorem 7.7.1:  $I(\Delta L_i; U_i | \tilde{Z}_i^+) \leq I(L_i, \bar{L}_i; U_i | \tilde{Z}_i^+) \leq I(F_i, \bar{F}_i; U_i | \tilde{Z}_i^+) \leq I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)$ . Additionally, notice that we can also apply the similar technique for the hypotheses-conditioned CMI, which should give the same results.

**Expressiveness of New CMI Notions Under Distribution-Free Setting** Previous works [199, 75, 76, 82, 92] have demonstrated that the CMI framework is expressive enough to establish connections with VC theory in the distribution-free learning setting. Here, we further illustrate the expressiveness of the sample-conditioned CMI discussed in this chapter.



**Theorem 7.7.2.** *Let  $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ , and let  $\mathcal{F} = \{f_w : \mathcal{X} \rightarrow \{0, 1\} | w \in \mathcal{W}\}$  be a functional hypothesis class with finite VC dimension  $d$ . Let  $n > d + 1$ , for any algorithm  $\mathcal{A}$ , we have  $\frac{1}{n} \sum_{i=1}^n \sqrt{I(F_i, \bar{F}_i; U_i | \tilde{Z}_i^+)} \leq \mathcal{O} \left( \sqrt{\frac{d}{n} \log \left( \frac{n}{d} \right)} \right)$ .*

Like the previous works, this bound matches the classic result of the uniform convergence bound [206]. Notice that the result could be extended to multi-class classification with finite Natarajan dimension [150] by proceeding similarly to [92, Thm. 8].

We invoke Theorem 7.7.2 to demonstrate that our new information-theoretic quantities have the same expressive power as standard CMI quantities. The expressiveness result for the (functional) hypotheses-conditioned CMI is expected to align with Theorem 7.7.2. This alignment is due to the equivalence between hypotheses-conditioned CMI and supsample-conditioned CMI, as discussed in Theorem 7.5.5.

## 7.8 Related Works and Additional Discussions

**Stability-Based Framework vs. Information-Theoretic Framework** Using stability methods to analyze generalization errors can be traced back to several seminal works, such as [181, 48, 49, 136, 109]. It is worth noting that stability arguments have proven particularly effective in analyzing the learnability of SCO problems [188], where traditional uniform convergence bounds may not be sufficient to explain the generalization behavior. The application of stability approaches has gained popularity for providing high-probability guarantees since the work of [24]. Recent advancements have further sharpened the convergence rates of high-probability generalization upper bounds for uniformly stable algorithms in a series of works [54, 55, 26, 112]. While information-theoretic generalization bounds are commonly used to analyze the in-expectation generalization, it is expected that combining them with the stability framework will yield sharper high-probability bounds.

Additionally, we note that in the realizable setting, where  $\mathcal{A}$  is an interpolating algorithm, information-theoretic generalization bounds exhibit greater power than stability-based bounds. For instance, in the case of the 0 – 1 loss, information-theoretic generalization bounds can achieve the known optimal minimax rates [75] and even provide exact characterizations of the generalization error [76, 217]. However, it should be noted that due to the inherent fitting-stability tradeoff property [186, Sec. 13.4], interpolating algorithms tend to be unstable, rendering stability arguments inapplicable. Given the prevalence of zero empirical risk in modern deep learning [238], it is natural to question whether

information-theoretic generalization bounds require the stability-based approach for analyzing non-convex (and potentially non-smooth and non-Lipschitz continuous) learning scenarios.

**Comparison with Standard CMI** While our IOMI quantity aligns with the previous work [29], the new CMI quantity  $I(\widehat{Z}_i; U_i | \widetilde{W}_i)$  (or equivalently  $I(W_i, \overline{W}_i; U_i | \widetilde{Z}_i^+)$ ) may not be directly comparable to the standard individual CMI,  $I(W; U_i | \widetilde{Z}_i)$ , in [179, 247]. Specifically, we have  $I(\widehat{Z}_i; U_i | \widetilde{W}_i) = H(U_i) - H(U_i | \widehat{Z}_i, \widetilde{W}_i)$  and  $I(W; U_i | \widetilde{Z}_i) = H(U_i) - H(U_i | W, \widetilde{Z}_i)$ . The relationship between  $H(U_i | \widehat{Z}_i, \widetilde{W}_i)$  and  $H(U_i | W, \widetilde{Z}_i)$  is not trivial. Exploring and quantitatively comparing these CMI measures would be an intriguing research direction.

**Leave-One-Out CMI** Our construction of  $\widetilde{W}$  bears resemblance to the leave-one-out (LOO) setting, where there are also  $n + 1$  hypotheses. It is worth noting that LOO-CMI has been recently proposed in concurrent works [76, 175]. In LOO-CMI, the supersample  $\widetilde{Z} = Z_{[n+1]}$  consists of  $n + 1$  instances, and  $U$  is an index uniformly drawn from  $[n + 1]$  to select one hold-out instance. Consequently,  $Z_U$  represents the testing data, while  $Z_{[n+1] \setminus U}$  serves as the training sample. In this context, LOO-CMI can be defined as  $I(W; U | \widetilde{Z})$ . Notice that this quantity still fails to explain the generalization in Example 7.1. The LOO setting is often associated with the stability-based framework [24], and it is expected that the  $n + 1$ -supersample induced  $\widetilde{W}$  could yield new CMI bounds that also contain the SCH stability notions. Nevertheless, in this paper, we do not adopt the LOO setting because in that case,  $H(U) = \log(n + 1)$ , the LOO-CMI bound is no longer upper bounded by a constant independent of  $n$  (note that the LOO-CMI bound for general setting in [76, Thm. 2.5] does not contain the  $1/\sqrt{n}$  factor).

**Stochastic Gradient Descent (SGD)** Since the influential work of [80], stability approaches have been widely employed to provide generalization guarantees for (sub)gradient-based optimization algorithms, such as SGD, under certain conditions like the convex-smooth-Lipschitz loss. More recently, [16] extended the results of [80] to the non-smooth loss function in the SCO setting.

In contrast, information-theoretic (weight/hypothesis-based) generalization bounds are typically used to analyze the noisy version of SGD, known as SGLD [168, 151, 77, 179, 210]. Directly analyzing SGD poses challenges because the returned hypothesis  $W$  contains a significant amount of information about  $S$  or  $Z_i$ , resulting in potentially large (even infinite) mutual information. The prevalent approach to applying information-theoretic

generalization bounds to SGD is by introducing noise [153, 213], but this has been shown to yield non-vanishing bounds in [78, Thm. 4].

The combination of information-theoretic generalization bounds with stability for analyzing the generalization of SGD presents a promising future direction. However, some potential difficulties may arise. For instance, if we continue to use the Gaussian noise perturbation method for the weight-based information-theoretic generalization bounds, we would need to characterize the stability property for the perturbed SGD, which might require techniques employed in [146]. Additionally, when combining stability notions with loss difference based CMI (or e-CMI/ $f$ -CMI) bounds, as they cannot be unrolled using the chain rule and data processing inequality as in the case of weight-based IOMI/CMI bounds, it may not be possible to bound such CMI terms using trajectory-based quantities. This raises doubts about the potential for obtaining more informative generalization bounds compared to the stability-based bounds themselves.

**Generalization Bounds beyond Mutual Information** In the information-theoretic literature, it is common to replace mutual information with alternative distributional measures, such as Wasserstein distance based bounds and total variation based bounds [180]. A promising future direction is to incorporate the stability property of algorithms into these bounds, as demonstrated in this chapter. It is worth noting that obtaining KL divergence-based bounds should be straightforward since they rely on the same foundational Lemma 2.4.1 as discussed in Chapter 7.

**Other Limitations** Our bounds, much like most information-theoretic generalization bounds, are derived under the assumption of independent and identically distributed (i.i.d.) training data, which is a common assumption in the field. Moreover, unless specific conditions are met, estimating stability notions can be challenging in general. For instance, in the case of non-convex learning tasks like deep learning, obtaining accurate estimations of stability notions can pose difficulties. Consequently, computing the bound based on these notions becomes a complex empirical task in practice.

## 7.9 Conclusion

In this chapter, we propose a novel construction of the hypothesis matrix and a new family of stability notions called sample-conditioned hypothesis stability. Leveraging these concepts, we derive sharper information-theoretic generalization bounds for stable learning

algorithms. Several promising avenues for future research include comparing our new CMI quantities with the standard CMI in a quantitative manner, analyzing the generalization of gradient-based optimization algorithms like SGD using our bounds, and establishing new high-probability generalization guarantees.

# Chapter 8

## Conclusion and Future Work

In this dissertation, we use an information-theoretic approach to analyze the generalization in machine learning systems.

First, our investigation yields a series of information-theoretic generalization bounds tailored for models trained with Stochastic Gradient Descent (SGD), with the aid of an auxiliary weight process. These bounds offer insights into the behavior of modern neural networks and inspire innovative regularization techniques.

Furthermore, we develop more information-theoretic bounds for models trained with SGD, incorporating the Stochastic Differential Equation (SDE) approximation of SGD. They are derived either from the trajectories of SGD under SDE modeling or based on the terminal state of SGD under the same modeling approach.

Beyond the realm of SGD, we also introduce new information-theoretic bounds for all black-box algorithms. These bounds provide alternative perspectives for understanding generalization, notably from a communication standpoint. Additionally, they either possess theoretical tightness or demonstrate empirical superiority over previous information-theoretic bounds within the same setting.

Shifting our focus to the domain adaptation scenario, we establish new generalization bounds for two distinct notions of generalization. Among these new bounds, the algorithm-dependent ones serve as valuable foundations for enhancing domain adaptation algorithms.

In the final segment of this dissertation, we propose a potential solution to address the limitations of existing information-theoretic bounds showcased in the literature. Specifically, we combine the stability-based analysis framework with the information-theoretic

analysis framework to derive novel generalization bounds capable of surmounting the aforementioned limitations.

The developments presented in this dissertation open up several promising avenues for future research:

### **Deep Learning Optimization and Information-Theoretic Generalization Bounds**

Chapters 3 and 4 emphasize the pivotal role of the training trajectory in determining model generalization. These findings, along with insights from [130], suggest that algorithm-specific optimization properties and termination criteria could be more influential than the choice of loss function or optimization solution. Exploring the relationship between optimization and generalization in deep learning, particularly considering factors like learning rate, batch size, and momentum, holds the potential for deriving more informative algorithm-dependent generalization bounds. Furthermore, it is also possible to apply the information-theoretic analysis framework developed in these sections to understand the generalization capabilities of more advanced deep neural networks, such as transformers [207], and to elucidate the fundamental factors influencing their training dynamics or the ultimate solutions leading to their superior performance.

**Heavy-Tailed SGD Noise** While Chapter 4 models SGD using the Itô SDE framework, recent research suggests that SGD gradient noise may exhibit heavy-tailed behavior, such as following an  $\alpha$ -stable distribution rather than a Gaussian one [194, 159, 195, 141, 73]. Investigating information-theoretic generalization bounds based on Lévy SDE under such conditions is a promising direction.

**Generalization Theory and Online Learning** In a recent study [135], a connection is drawn between the online learning setting and the statistical learning setting, leading to the derivation of generalization bounds from an online learning perspective. Notably, their framework has the capability to recover some existing generalization bounds, including PAC-Bayesian and information-theoretic bounds. We believe that applying such a framework to developing novel information-theoretic generalization bounds, particularly under the setting in Chapter 5, holds significant promise.

**$f$ -divergence Guided Domain Adaptation Framework** Chapter 6 provides theoretical support for minimizing the KL divergence between source and target domains in the representation space to enhance generalization on unseen target data. Building upon

this foundation, it is natural to extend the theoretical framework to a broader family of divergences known as  $f$ -divergences, with KL divergence being just one example. This extension could draw upon recent advances in variational representation for  $f$ -divergences, as explored in [218], serving as a promising starting point.

These research directions not only promise to deepen our understanding of the topics explored in this dissertation but also have the potential to advance the field of machine learning and information theory.

# References

- [1] Alessandro Achille, Giovanni Paolini, and Stefano Soatto. Where is the information in a deep neural network? *arXiv preprint arXiv:1905.12213*, 2019.
- [2] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.
- [3] David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pages 66–75. PMLR, 2021.
- [4] Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. In *International Conference on Machine Learning*, pages 115–124. PMLR, 2020.
- [5] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.
- [6] Pierre Alquier. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.
- [7] Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. An information-theoretical approach to semi-supervised learning under covariate-shift. In *International Conference on Artificial Intelligence and Statistics*, pages 7433–7449. PMLR, 2022.
- [8] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*. PMLR, 2019.



- [9] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.
- [10] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.
- [11] Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.
- [12] Arindam Banerjee, Tiancong Chen, Xinyan Li, and Yingxue Zhou. Stability based generalization bounds for exponential family langevin dynamics. In *International Conference on Machine Learning*, pages 1412–1449. PMLR, 2022.
- [13] David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.
- [14] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 2017.
- [15] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.
- [16] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.
- [17] Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*. PMLR, 2018.
- [18] Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.
- [19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

- [20] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.
- [21] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.
- [22] Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.
- [23] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.
- [24] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
- [25] Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020.
- [26] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.
- [27] Jean Bretagnolle and Catherine Huber. Estimation des densités: risque mini-max. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.
- [28] Yuheng Bu, Gholamali Aminian, Laura Toni, Gregory W Wornell, and Miguel Rodrigues. Characterizing and understanding the generalization error of transfer learning with gibbs algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 8673–8699. PMLR, 2022.
- [29] Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591. IEEE, 2019.
- [30] Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gürbüzbalaban, and Umut Şimşekli. Asymmetric heavy tails and implicit bias in gaussian noise injections. *arXiv preprint arXiv:2102.07006*, 2021.

- [31] Alexander Camuto, Matthew Willetts, Umut Simsekli, Stephen J. Roberts, and Chris C. Holmes. Explicit regularisation in gaussian noise injections. In *Advances in Neural Information Processing Systems*, 2020.
- [32] Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.
- [33] O Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *Vol. 56. Lecture Notes - Monograph Series. Institute of Mathematical Statistics*, 2007.
- [34] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations*. OpenReview.net, 2017.
- [35] Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. *Advances in Neural Information Processing Systems*, 34, 2021.
- [36] Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.
- [37] Eugenio Clerico, Amitis Shidani, George Deligiannidis, and Arnaud Doucet. Chained generalisation bounds. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4212–4257. PMLR, 02–05 Jul 2022.
- [38] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.
- [39] R Dennis Cook and Sanford Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.
- [40] Corinna Cortes, Spencer Greenberg, and Mehryar Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85(1):45–70, 2019.
- [41] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

- [42] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.
- [43] Gavin E. Crooks. Inequalities between the jenson-shannon and jeffreys divergences. In *Tech. Note 004*, 2008.
- [44] Villani Cédric. *Optimal Transport: Old and New (Grundlehren der mathematischen Wissenschaften, 338)*. Springer, 2008.
- [45] Felix Dangel, Frederik Kunstner, and Philipp Hennig. Backpack: Packing more into backprop. In *8th International Conference on Learning Representations*. OpenReview.net, 2020.
- [46] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.
- [47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.
- [48] Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.
- [49] Luc Devroye and Terry Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.
- [50] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.
- [51] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.
- [52] Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbling. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

- [53] Fartash Faghri, David Duvenaud, David J Fleet, and Jimmy Ba. A study of gradient variance in deep learning. *arXiv preprint arXiv:2007.04532*, 2020.
- [54] Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.
- [55] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019.
- [56] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.
- [57] Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. *Advances in Neural Information Processing Systems*, 32, 2019.
- [58] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.
- [59] Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer Berlin, Heidelberg, 2012.
- [60] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.
- [61] Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. In *International Conference on Learning Representations*, 2022.
- [62] Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in  $r^d$ . *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.
- [63] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. Pac-bayes and domain adaptation. *Neurocomputing*, 379:379–397, 2020.

- [64] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 315–323, 2011.
- [65] Ziv Goldfeld and Kristjan Greenewald. Sliced mutual information: A scalable measure of statistical dependence. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [66] Ziv Goldfeld, Kristjan Greenewald, Theshani Nuradha, and Galen Reeves.  $\mathbb{S}^k$ -sliced mutual information: A quantitative study of scalability with dimension. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [67] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR*, 2015.
- [68] Peter Grunwald, Thomas Steinke, and Lydia Zakyntinou. Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes. In *Conference on Learning Theory*. PMLR, 2021.
- [69] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.
- [70] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.
- [71] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE transactions on information theory*, 51(4):1261–1282, 2005.
- [72] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.
- [73] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*. PMLR, 2021.
- [74] Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33:16457–16467, 2020.

- [75] Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34:26370–26381, 2021.
- [76] Mahdi Haghifam, Shay Moran, Daniel M Roy, and Gintare Karolina Dziugaite. Understanding generalization via leave-one-out conditional mutual information. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2487–2492. IEEE, 2022.
- [77] Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 2020.
- [78] Mahdi Haghifam, Borja Rodríguez-Gálvez, Ragnar Thobaben, Mikael Skoglund, Daniel M Roy, and Gintare Karolina Dziugaite. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In *International Conference on Algorithmic Learning Theory*, pages 663–706. PMLR, 2023.
- [79] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.
- [80] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.
- [81] Hrayr Harutyunyan, Alessandro Achille, Giovanni Paolini, Orchid Majumder, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Estimating informativeness of samples with smooth unique information. In *International Conference on Learning Representations*, 2021.
- [82] Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, 2021.
- [83] Hrayr Harutyunyan, Kyle Reing, Greg Ver Steeg, and Aram Galstyan. Improving generalization by controlling label-noise information in neural network weights. In *International Conference on Machine Learning*, pages 4071–4081. PMLR, 2020.

- [84] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems 32*, pages 1141–1150, 2019.
- [85] Haiyun He, Hanshu Yan, and Vincent YF Tan. Information-theoretic generalization bounds for iterative semi-supervised learning. *arXiv preprint arXiv:2110.00926*, 2021.
- [86] Haiyun He, Hanshu Yan, and Vincent YF Tan. Information-theoretic characterization of the generalization error for iterative semi-supervised learning. *Journal of Machine Learning Research*, 23:1–52, 2022.
- [87] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [88] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645, 2016.
- [89] Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 2020.
- [90] Fredrik Hellström and Giuseppe Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 952–957. IEEE, 2021.
- [91] Fredrik Hellström and Giuseppe Durisi. Evaluated CMI bounds for meta learning: Tightness and expressiveness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [92] Fredrik Hellström and Giuseppe Durisi. A new family of generalization bounds using samplewise evaluated CMI. In *Advances in Neural Information Processing Systems*, 2022.
- [93] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.
- [94] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.



- [95] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.
- [96] Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations*, 2020.
- [97] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.
- [98] Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th International Conference on Machine Learning, ICML, 2020*.
- [99] Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*. PMLR, 2021.
- [100] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.
- [101] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.
- [102] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.
- [103] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479. IEEE, 2017.
- [104] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

- [105] Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic bounds on transfer generalization gap based on jensen-shannon divergence. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1461–1465. IEEE, 2021.
- [106] Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126, 2021.
- [107] Sharu Theresa Jose, Osvaldo Simeone, and Giuseppe Durisi. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization. *IEEE Transactions on Information Theory*, 68(1):474–501, 2021.
- [108] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.
- [109] Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 152–162, 1997.
- [110] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations*. OpenReview.net, 2017.
- [111] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [112] Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate  $o(1/n)$ . *Advances in Neural Information Processing Systems*, 34:5065–5076, 2021.
- [113] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.
- [114] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [115] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

- [116] Brett W Larsen, Stanislav Fort, Nic Becker, and Surya Ganguli. How many degrees of freedom do we need to train deep networks: a loss landscape perspective. In *International Conference on Learning Representations*, 2022.
- [117] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- [118] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.
- [119] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.
- [120] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.
- [121] Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.
- [122] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*. PMLR, 2017.
- [123] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.
- [124] Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sgd for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 2020.
- [125] Zhiyuan Li, Sathika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 2021.
- [126] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

- [127] Nick Littlestone. Relating data compression and learnability. *Technical report*, 1986.
- [128] Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, pages 7045–7056. PMLR, 2021.
- [129] Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017.
- [130] Zixuan Liu, Ziqiao Wang, Hongyu Guo, and Yongyi Mao. Over-training with mixup may hurt generalization. In *The Eleventh International Conference on Learning Representations*, 2023.
- [131] Roi Livni. Information theoretic lower bounds for information theoretic upper bounds. *arXiv preprint arXiv:2302.04925*, 2023.
- [132] Roi Livni and Shay Moran. A limitation of the pac-bayes framework. *Advances in Neural Information Processing Systems*, 33:20543–20553, 2020.
- [133] Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2935–2944, 2017.
- [134] Gábor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In *Conference on Learning Theory*, pages 3524–3546. PMLR, 2022.
- [135] Gábor Lugosi and Gergely Neu. Online-to-pac conversions: Generalization bounds via regret analysis. *arXiv preprint arXiv:2305.19674*, 2023.
- [136] Gábor Lugosi and Mirosław Pawlak. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Transactions on Information Theory*, 40(2):475–481, 1994.
- [137] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18:1–35, 2017.
- [138] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *The 22nd Conference on Learning Theory*, 2009.

- [139] Mohammad Saeed Masiha, Amin Gohari, Mohammad Hossein Yassaee, and Mohammad Reza Aref. Learning under distribution mismatch and model misspecification. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2912–2917. IEEE, 2021.
- [140] Colin McDiarmid. *Concentration*, pages 195–248. Springer Berlin Heidelberg, 1998.
- [141] Qi Meng, Shiqi Gong, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Dynamic of stochastic gradient descent with state-dependent noise. *arXiv preprint arXiv:2006.13719*, 2020.
- [142] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. In *International Conference on Learning Representations*, 2018.
- [143] Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. Pac-bayes un-expected bernstein inequality. *Advances in Neural Information Processing Systems*, 32, 2019.
- [144] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2017.
- [145] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [146] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.
- [147] Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019.
- [148] Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [149] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.
- [150] Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

- [151] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 2019.
- [152] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [153] Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*. PMLR, 2021.
- [154] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 2017.
- [155] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.
- [156] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2018.
- [157] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.
- [158] A. Tuan Nguyen, Toan Tran, Yarin Gal, Philip Torr, and Atilim Gunes Baydin. KL guided domain adaptation. In *International Conference on Learning Representations*, 2022.
- [159] Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *Advances in neural information processing systems*, 2019.
- [160] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

- [161] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.
- [162] Daniel P Palomar and Sergio Verdú. Lautum information. *IEEE transactions on information theory*, 54(3):964–975, 2008.
- [163] Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli. Non-gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.
- [164] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.
- [165] Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.
- [166] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [167] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.
- [168] Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018.
- [169] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.
- [170] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for 6.441 (MIT), ECE 563 (UIUC), STAT 364 (Yale), 2019.*, 2019.
- [171] Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9(129-145):20, 1935.

- [172] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.
- [173] Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30. IEEE, 2016.
- [174] Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications and coding. *Foundations and Trends in Communications and Information Theory; NOW Publishers: Boston, MA, USA*, 2018.
- [175] Mohamad Rida Rammal, Alessandro Achille, Aditya Golatkar, Suhas Diggavi, and Stefano Soatto. On leave-one-out conditional mutual information for generalization. In *Advances in Neural Information Processing Systems*, 2022.
- [176] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younes Benani. A survey on domain adaptation theory. *arXiv preprint arXiv:2004.11829*, 2020.
- [177] Arezou Rezazadeh, Sharu Theresa Jose, Giuseppe Durisi, and Osvaldo Simeone. Conditional mutual information-based generalization bound for meta learning. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1176–1181. IEEE, 2021.
- [178] Omar Rivasplata, Emilio Parrado-Hernández, John S Shawe-Taylor, Shiliang Sun, and Csaba Szepesvári. Pac-bayes bounds for stable algorithms with instance-dependent priors. *Advances in Neural Information Processing Systems*, 31, 2018.
- [179] Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. In *2020 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2021.
- [180] Borja Rodríguez Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter expected generalization error bounds via wasserstein distance. *Advances in Neural Information Processing Systems*, 34, 2021.
- [181] William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.



- [182] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*, pages 2663–2671, 2012.
- [183] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*. PMLR, 2016.
- [184] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.
- [185] Yevgeny Seldin, François Laviolette, Nicolo Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.
- [186] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [187] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009.
- [188] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.
- [189] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [190] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [191] Maohao Shen, Yuheng Bu, and Gregory Wornell. On the benefits of selectivity in pseudo-labeling for unsupervised multi-source-free domain adaptation. *arXiv preprint arXiv:2202.00796*, 2022.
- [192] Changjian Shui, Qi Chen, Jun Wen, Fan Zhou, Christian Gagné, and Boyu Wang. Beyond  $\mathcal{H}$ -divergence: Domain adaptation theory with jensen-shannon divergence. *arXiv preprint arXiv:2007.15567*, 2020.

- [193] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015.
- [194] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*. PMLR, 2019.
- [195] Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.
- [196] Sidak Pal Singh, Aurelien Lucchi, Thomas Hofmann, and Bernhard Schölkopf. Phenomenology of double descent in finite-width neural networks. In *International Conference on Learning Representations*, 2022.
- [197] Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.
- [198] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [199] Thomas Steinke and Lydia Zakyntinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*. PMLR, 2020.
- [200] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.
- [201] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2818–2826, 2016.
- [202] Kei Takeuchi. The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science*, 153:12–18, 1976.
- [203] Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature

- and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR, 2020.
- [204] Ilya O Tolstikhin and Yevgeny Seldin. Pac-bayes-empirical-bernstein inequality. *Advances in Neural Information Processing Systems*, 26, 2013.
  - [205] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
  - [206] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.
  - [207] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
  - [208] Bohan Wang, Huishuai Zhang, Jieyu Zhang, Qi Meng, Wei Chen, and Tie-Yan Liu. Optimizing information-theoretical generalization bound via anisotropic noise of sgld. *Advances in Neural Information Processing Systems*, 2021.
  - [209] Boyu Wang, Jorge Mendez, Mingbo Cai, and Eric Eaton. Transfer learning via minimizing the performance gap between domains. *Advances in Neural Information Processing Systems*, 32, 2019.
  - [210] Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of sgld using properties of gaussian channels. *Advances in Neural Information Processing Systems*, 34:24222–24234, 2021.
  - [211] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021.
  - [212] Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. PAC-bayes information bottleneck. In *International Conference on Learning Representations*, 2022.
  - [213] Ziqiao Wang and Yongyi Mao. On the generalization of models trained with SGD: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2022.
  - [214] Ziqiao Wang and Yongyi Mao. Two facets of sde under an information-theoretic lens: Generalization of sgld via training trajectories and via terminal states. *arXiv preprint arXiv:2211.10691*, 2022.

- [215] Ziqiao Wang and Yongyi Mao. Information-theoretic analysis of unsupervised domain adaptation. In *International Conference on Learning Representations*, 2023.
- [216] Ziqiao Wang and Yongyi Mao. Sample-conditioned hypothesis stability sharpens information-theoretic generalization bounds. In *Advances in Neural Information Processing Systems*, 2023.
- [217] Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. *arXiv preprint arXiv:2302.02432*, 2023.
- [218] Ziqiao Wang and Yongyi Mao. On f-divergence principled domain adaptation: An improved framework. *arXiv preprint arXiv:2402.01887*, 2024.
- [219] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International Conference on Machine Learning*, pages 10181–10192. PMLR, 2020.
- [220] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.
- [221] Yeming Wen, Kevin Luk, Maxime Gazeau, Guodong Zhang, Harris Chan, and Jimmy Ba. An empirical study of stochastic gradient descent with structured covariance noise. In *International Conference on Artificial Intelligence and Statistics*, pages 3621–3631. PMLR, 2020.
- [222] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.
- [223] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*. PMLR, 2020.
- [224] Lei Wu, Chao Ma, and E Weinan. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 2018:8279–8288, 2018.
- [225] Xuetong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. Information-theoretic analysis for transfer learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2819–2824. IEEE, 2020.

- [226] Xu Tong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. Fast rate generalization error bounds: Variations on a theme. In *2022 IEEE Information Theory Workshop (ITW)*, pages 43–48. IEEE, 2022.
- [227] Xu Tong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. On the tightness of information-theoretic bounds on generalization error of learning algorithms. *arXiv preprint arXiv:2303.14658*, 2023.
- [228] Zeke Xie, Fengxiang He, Shaopeng Fu, Issei Sato, Dacheng Tao, and Masashi Sugiyama. Artificial neural variability for deep learning: on overfitting, noise memorization, and catastrophic forgetting. *Neural computation*, 33(8):2163–2192, 2021.
- [229] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.
- [230] Zeke Xie, Li Yuan, Zhanxing Zhu, and Masashi Sugiyama. Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. In *International Conference on Machine Learning*. PMLR, 2021.
- [231] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.
- [232] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017.
- [233] Jun Yang, Shengyang Sun, and Daniel M Roy. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. *Advances in Neural Information Processing Systems*, 32, 2019.
- [234] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.
- [235] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 581–590. IEEE, 2020.
- [236] Raymond W Yeung. A new outlook on shannon’s information measures. *IEEE transactions on information theory*, 37(3):466–474, 1991.

- [237] Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. *arXiv preprint arXiv:2202.05834*, 2022.
- [238] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.
- [239] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR*, 2018.
- [240] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019.
- [241] Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.
- [242] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019.
- [243] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.
- [244] Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8156–8165, 2021.
- [245] Nikita Zhivotovskiy and Steve Hanneke. Localization of vc classes: Beyond local rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.
- [246] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv e-prints*, pages arXiv–2103, 2021.
- [247] Ruida Zhou, Chao Tian, and Tie Liu. Individually conditional individual mutual information bound on generalization error. *IEEE Transactions on Information Theory*, 68(5):3304–3316, 2022.

- [248] Ruida Zhou, Chao Tian, and Tie Liu. Stochastic chaining and strengthened information-theoretic generalization bounds. In *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022.
- [249] Ruida Zhou, Chao Tian, and Tie Liu. Exactly tight information-theoretic generalization error bound for the quadratic gaussian problem. *arXiv preprint arXiv:2305.00876*, 2023.
- [250] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*. PMLR, 2019.
- [251] Liu Ziyin, Botao Li, James B Simon, and Masahito Ueda. SGD can converge to local maxima. In *International Conference on Learning Representations*, 2022.
- [252] Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022.

# APPENDICES



# Appendix A

## Appendix of Chapter 3

### A.1 HWI Inequality: Proof of Lemma 3.4.1

The proof given here is a simple extension of the proof of [174, Lemma 3.4.2], which is a special instance of the (weak) HWI inequality.

*Proof.*

$$D_{\text{KL}}(P_{X+\sqrt{t}N}||P_{Y+\sqrt{t'}N}) \leq D_{\text{KL}}(P_{X,Y,X+\sqrt{t}N}||P_{X,Y,Y+\sqrt{t'}N}) \quad (\text{A.1})$$

$$\begin{aligned} &= \mathbb{E}_{X,Y} \left[ D_{\text{KL}}(P_{X+\sqrt{t}N|X,Y}||P_{Y+\sqrt{t'}N|X,Y}) \right] \\ &= \mathbb{E}_{X,Y} [D_{\text{KL}}(\mathcal{N}(X, t\mathbf{I}_d)||\mathcal{N}(Y, t'\mathbf{I}_d))] \end{aligned} \quad (\text{A.2})$$

$$= \frac{1}{2t'} \mathbb{E}_{X,Y} [\|X - Y\|^2] + \frac{d}{2} \left( \log \frac{t'}{t} + \frac{t}{t'} - 1 \right), \quad (\text{A.3})$$

where Eq. (A.1) is by the chain rule of the KL divergence, Eq. (A.2) holds since  $N$  is independent of  $(X, Y)$  and Eq. (A.3) is by the equality

$$D_{\text{KL}}(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2} \left[ \log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^{\mathbf{T}}\Sigma_2^{-1}(\mu_2 - \mu_1) \right].$$

Eq. (A.3) holds for any joint distribution of  $(X, Y)$ . If we let  $t' = t$ , by the definition of Wasserstein distance, we have

$$D_{\text{KL}}(P_{X+\sqrt{t}N}||P_{Y+\sqrt{t}N}) \leq \frac{1}{2t} \mathbb{W}_2^2(P_X, P_Y).$$

This recovers the vector version of [174, Lemma 3.4.2]. □

## A.2 A Key Lemma: Lemma 3.4.2

*Proof.* We give two proofs of this lemma, the first proof is inspired by the techniques used in [210] and [71], and the second proof is similar to [168].

1. Let  $Z = Y + \Delta$ . By the properties of mutual information,

$$I(f(Z, X) + \sigma N; X|Y = y) = I(f(Z, X) - \Omega + \sigma N; X|Y = y).$$

Let  $g(Z, X) \triangleq f(Z, X) - \Omega$ . Let  $P_{N'} = \mathcal{N}(0, \sigma'^2 \mathbf{I}_d)$  for any  $\sigma' > 0$  and  $N'$  is independent of other random variables for a fixed  $y$ . Then, for  $Y = y$ ,

$$I(g(Z, X) + \sigma N; X|Y = y) \tag{A.4}$$

$$\begin{aligned} &= \mathbb{E}_X [\text{D}_{\text{KL}}(P_{g(Z, x) + \sigma N|X=x, Y=y} \| P_{g(Z, X) + \sigma N|Y=y})] \\ &= \mathbb{E}_X [\text{D}_{\text{KL}}(P_{g(Z, x) + \sigma N|X=x, Y=y} \| P_{N'|Y=y}) - \text{D}_{\text{KL}}(P_{g(Z, X) + \sigma N|Y=y} \| P_{N'|Y=y})] \end{aligned} \tag{A.5}$$

$$\leq \inf_{\sigma' > 0} \mathbb{E}_X [\text{D}_{\text{KL}}(P_{g(Z, x) + \sigma N|X=x, Y=y} \| P_{\sigma' N|Y=y})] \tag{A.6}$$

$$\begin{aligned} &= \inf_{\sigma' > 0} \frac{1}{2\sigma'^2} \mathbb{E}_X [\mathbb{E}_\Delta [||g(y + \Delta, x)||^2 | X = x, Y = y]] + \frac{d}{2} (\log \frac{\sigma'^2}{\sigma^2} + \frac{\sigma^2}{\sigma'^2} - 1) \tag{A.7} \\ &= \frac{d}{2} \log \left( \frac{\mathbb{E}_{X, \Delta} [||g(y + \Delta, X)||^2 | Y = y]}{d\sigma^2} + 1 \right), \end{aligned}$$

where Eq. (A.5) is by the golden formula (cf. Lemma 2.4.10), Eq. (A.6) is by the non-negativity of relative entropy, Eq. (A.7) is due to Lemma 3.4.1 and the last equality holds when the optimal

$$\sigma'^2 = \frac{\mathbb{E}_{X, \Delta} [||g(y + \Delta, X)||^2 | Y = y]}{d} + \sigma^2$$

is achieved.

Thus,

$$\begin{aligned} I(f(Z, X) + N; X|Y) &= \mathbb{E}_Y [I(f(Z, X) + N; X|Y = y)] \\ &\leq \mathbb{E}_Y \left[ \frac{d}{2} \log \left( \frac{\mathbb{E}_{X, \Delta} [||g(y + \Delta, X)||^2 | Y = y]}{d\sigma^2} + 1 \right) \right] \\ &= \mathbb{E}_Y \left[ \frac{d}{2} \log \left( \frac{\mathbb{E}_{X, \Delta} [||f(Z, X) - \Omega||^2 | Y = y]}{d\sigma^2} + 1 \right) \right], \end{aligned}$$

which completes the proof.

2. Alternatively, let  $h(\cdot)$  denote the differential entropy, then

$$\begin{aligned}
I(f(Z, X) + \sigma N; X|Y = y) &= I(g(Z, X) + \sigma N; X|Y = y) \\
&= h(g(Z, X) + \sigma N|Y = y) - h(g(Z, X) + \sigma N|X, Y = y) \\
&\leq h(g(Z, X) + \sigma N|Y = y) - h(g(Z, X) + \sigma N|\Delta, X, Y = y) \\
&\quad \quad \quad (\text{A.8}) \\
&= h(g(Z, X) + \sigma N|Y = y) - h(\sigma N) \\
&= h(g(Z, X) + \sigma N|Y = y) - \frac{d}{2} \log 2\pi e \sigma^2,
\end{aligned}$$

where Eq. (A.8) is due to the fact that conditioning reduces entropy.

Notice that

$$\mathbb{E}_{\Delta, X, N} [||g(y + \Delta, X) + \sigma N||^2 | Y = y] = \mathbb{E}_{\Delta, X} [||g(y + \Delta, X)||^2 | Y = y] + d\sigma^2.$$

Since the Gaussian distribution maximizes the entropy over all distributions with the same variance [41], we have

$$h(g(Z, X) + \sigma N|Y = y) \leq \frac{d}{2} \log 2\pi e \frac{\mathbb{E}_{\Delta, X} [||g(y + \Delta, X)||^2 | Y = y] + d\sigma^2}{d}.$$

Therefore,

$$\begin{aligned}
I(f(Z, X) + \sigma N; X|Y = y) &\leq \frac{d}{2} \log 2\pi e \frac{\mathbb{E}_{\Delta, X} [||g(y + \Delta, X)||^2 | Y = y] + d\sigma^2}{d} - \frac{d}{2} \log 2\pi e \sigma^2 \\
&= \frac{d}{2} \log \left( \frac{\mathbb{E}_{\Delta, X} [||g(y + \Delta, X)||^2 | Y = y]}{d\sigma^2} + 1 \right).
\end{aligned}$$

The remaining part is straightforward and is the same with the previous proof.

□

### A.3 Mutual Information Bounds for SGD: Theorem 3.4.2

In this section, we provide a proof for Theorem 3.4.2, while elaborating on why the local gradient sensitivity term in Lemma 3.3.1 can be removed. The key ingredient that makes this possible is a construction given in the proof of Lemma 3.4.3.

### A.3.1 Lemma A.3.1

We first give the following useful lemma.

**Lemma A.3.1.** *The mutual information  $I(\widetilde{W}_T; S) \leq \sum_{t=1}^T I(-\lambda_t g(W_{t-1}, B_t) + N_t; S | \widetilde{W}_{t-1})$ .*

*Proof.* The mutual information between the final output of SGD and the training sample can be upper bounded by the mutual information between the full trajectories and the training sample, which is shown below.

$$\begin{aligned} & I(\widetilde{W}_T; S) \\ &= I(\widetilde{W}_{T-1} - \lambda_T g(W_{T-1}, B_T) + N_T; S) \\ &\leq I(\widetilde{W}_{T-1}, -\lambda_T g(W_{T-1}, B_T) + N_T; S) \end{aligned} \tag{A.9}$$

$$= I(\widetilde{W}_{T-1}; S) + I(-\lambda_T g(W_{T-1}, B_T) + N_T; S | \widetilde{W}_{T-1}) \tag{A.10}$$

$$\leq I(\widetilde{W}_{T-2}; S) + I(-\lambda_{T-1} g(W_{T-2}, B_{T-1}) + N_{T-1}; S | \widetilde{W}_{T-2}) + I(-\lambda_T g(W_{T-1}, B_T) + N_T; S | \widetilde{W}_{T-1}) \tag{A.11}$$

⋮

$$\leq I(\widetilde{W}_0; S) + \sum_{t=1}^T I(-\lambda_t g(W_{t-1}, B_t) + N_t; S | \widetilde{W}_{t-1}), \tag{A.12}$$

$$= \sum_{t=1}^T I(-\lambda_t g(W_{t-1}, B_t) + N_t; S | \widetilde{W}_{t-1}), \tag{A.13}$$

where Eq. (A.9) is by  $I(f(X, Y); Z) \leq I(X, Y; Z)$  (i.e.  $Z - (X, Y) - f(X, Y)$  forms a Markov chain and then use the data processing inequality), Eq. (A.10) is by the chain rule of mutual information, Eq. (A.11) is by applying the similar procedure (namely, Eq. (A.9-A.10)) to  $I(\widetilde{W}_{T-1}; S)$ , Eq. (A.12) is by doing these steps recursively and Eq. (A.13) is due to the fact that  $\widetilde{W}_0$  is independent of  $S$  (i.e.  $I(\widetilde{W}_0; S) = 0$ ). This completes the proof.  $\square$

### A.3.2 Corollary A.3.1: Recovering Lemma 3.3.1

We can recover the result of Lemma 3.3.1 in [153] from Theorem 3.4.1, which is re-stated in the following corollary.

**Corollary A.3.1.** *The generalization error of SGD is upper bounded by*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{2R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}} \left[ 3\Psi(W_{t-1}) + 2\tilde{\mathbb{V}}_t(W_{t-1}) \right]} + |\mathbb{E}[\gamma(W_T, S) - \gamma(W_T, S')]|.$$

*Proof.* Let  $\tilde{e} = \mathbb{E}_Z [g(\tilde{W}_{t-1}, Z)]$ . From the first term in Theorem 3.4.1, we have

$$\begin{aligned} & \sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \mathbb{E}_{\tilde{W}_{t-1}} \left[ \log \left( \frac{\lambda_t^2 \mathbb{E}_{S, \Delta_{t-1}} \left[ \left\| g(W_{t-1}, B_t) - \mathbb{E}_{Z, \Delta_{t-1}} [g(W_{t-1}, Z)] \right\|^2 \right]}{d\sigma_t^2} + 1 \right) \right]} \\ & \leq \sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \log \left( \frac{\lambda_t^2 \mathbb{E}_{\tilde{W}_{t-1}, \Delta_{t-1}, S} \left[ \left\| g(W_{t-1}, B_t) - \mathbb{E}_{Z, \Delta_{t-1}} [g(\tilde{W}_{t-1} - \Delta_{t-1}, Z)] \right\|^2 \right]}{d\sigma_t^2} + 1 \right)} \end{aligned} \quad (\text{A.14})$$

$$\leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{\tilde{W}_{t-1}, \Delta_{t-1}, S} \left[ \left\| g(W_{t-1}, B_t) - \mathbb{E}_{Z, \Delta_{t-1}} [g(\tilde{W}_{t-1} - \Delta_{t-1}, Z)] \right\|^2 \right]} \quad (\text{A.15})$$

$$\begin{aligned} & = \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}, S} \left[ \left\| g(W_{t-1}, B_t) - \tilde{e} + \tilde{e} - \mathbb{E}_{Z, \Delta_{t-1}} [g(\tilde{W}_{t-1} - \Delta_{t-1}, Z)] \right\|^2 \right]} \\ & \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}, S} \left[ 2 \left\| g(W_{t-1}, B_t) - \tilde{e} \right\|^2 + 2 \left\| \tilde{e} - \mathbb{E}_{Z, \Delta_{t-1}} [g(\tilde{W}_{t-1} - \Delta_{t-1}, Z)] \right\|^2 \right]}, \end{aligned} \quad (\text{A.16})$$

where Eq. (A.14) is by Jensen's inequality (i.e. the concavity of logarithm), Eq. (A.15) is by  $\log(x+1) \leq x$ , and Eq. (A.16) is by  $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ .

Again by Jensen's inequality (i.e. the convexity of squared norm), we can further upper bound Eq. (A.16) by

$$\sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}, S} \left[ 2 \left\| g(W_{t-1}, B_t) - \tilde{e} \right\|^2 + 2 \mathbb{E}_{\Delta_{t-1}} \left[ \left\| \tilde{e} - \mathbb{E}_Z [g(\tilde{W}_{t-1} - \Delta_{t-1}, Z)] \right\|^2 \right] \right]}. \quad (\text{A.17})$$

Let  $e = \mathbb{E}_Z [g(W_{t-1}, Z)]$ , then

$$\begin{aligned}
& \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}, S} \left[ 2 \|g(W_{t-1}, B_t) - \tilde{e}\|^2 + 2 \mathbb{E}_{\Delta_{t-1}} \left[ \left\| \tilde{e} - \mathbb{E}_Z \left[ g(\widetilde{W}_{t-1} - \Delta_{t-1}, Z) \right] \right\|^2 \right] \right]} \\
&= \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}, S} \left[ 2 \|g(W_{t-1}, B_t) - e + e - \tilde{e}\|^2 + 2 \Psi(W_{t-1}) \right]} \\
&\leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}, S} \left[ 4 \|g(W_{t-1}, B_t) - e\|^2 + 4 \|e - \tilde{e}\|^2 + 2 \Psi(W_{t-1}) \right]} \tag{A.18} \\
&= \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}, S} \left[ 4 \|g(W_{t-1}, B_t) - e\|^2 + 6 \Psi(W_{t-1}) \right]} \\
&= \sqrt{\frac{2R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}} \left[ 3 \Psi(W_{t-1}) + 2 \widetilde{\mathbb{V}}_t(W_{t-1}) \right]},
\end{aligned}$$

where A.18 is by  $\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ .

This completes the proof.  $\square$

### A.3.3 Proof of Lemma 3.4.3

*Proof.* We first notice that

$$I(G_t + N_t; S | \widetilde{W}_{t-1}) = I(-\lambda_t g(W_{t-1}, B_t) + \sigma_t N; S | \widetilde{W}_{t-1}),$$

where  $N \sim \mathcal{N}(0, \mathbf{I}_d)$ .

Then let  $X = S$ ,  $Y = \widetilde{W}_{t-1}$  and  $\Omega = \mathbb{E}_{W_{t-1}, Z} [\nabla \ell(W_{t-1}, Z)]$ , by applying Lemma 3.4.2, we have

$$\begin{aligned}
& I(-\lambda_t g(W_{t-1}, B_t) + \sigma_t N; S | \widetilde{W}_{t-1}) \\
& \leq \frac{d}{2} \mathbb{E}_{\widetilde{W}_{t-1}} \left[ \log \left( \frac{\lambda_t^2 \mathbb{E}_{S, \Delta_{t-1}} [\|g(W_{t-1}, B_t) - \mathbb{E}_{W_{t-1}, Z} [\nabla_w \ell(W_{t-1}, Z)]\|^2]}{d \sigma_t^2} + 1 \right) \right] \\
& \leq \frac{d}{2} \log \left( \frac{\lambda_t^2 \mathbb{E}_{W_{t-1}} [\mathbb{V}_t(W_{t-1})]}{d \sigma_t^2} + 1 \right),
\end{aligned}$$

where the second inequality is by Jensen's inequality.

This completes the proof.  $\square$

### A.3.4 Proof of Theorem 3.4.2

*Proof.* Applying Lemma A.3.1 and Lemma 3.4.3 and putting everything together, we have

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \log \left( \frac{\lambda_t^2 \mathbb{E}_{W_{t-1}} [\mathbb{V}_t(W_{t-1})]}{d\sigma_t^2} + 1 \right)} + |\mathbb{E} [\gamma(W_T, S) - \gamma(W_T, S')]|.$$

Next, to handle the mismatch between the outputs of perturbed SGD and SGD, we apply Taylor expansion around  $\Delta_T = \vec{0}$ ,

$$\begin{aligned} \mathbb{E}_{W_T, S, \Delta_T} [L_S(W_T + \Delta_T) - L_S(W_T)] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W_T, Z_i, \Delta_T} [\ell(W_T + \Delta_T, Z_i) - \ell(W_T, Z_i)] \\ &\approx \mathbb{E}_{W_T, Z, \Delta_T} \left[ \langle \nabla_w \ell(W_T, Z), \Delta_T \rangle + \frac{1}{2} \Delta_T^T \mathbf{H}_{W_T}(Z) \Delta_T \right] \\ &= \mathbb{E}_{W_T, Z, \Delta_T} \left[ \frac{1}{2} \Delta_T^T \mathbf{H}_{W_T}(Z) \Delta_T \right] \end{aligned} \quad (\text{A.19})$$

$$\begin{aligned} &= \frac{1}{2} \langle \mathbb{E}_{W_T, Z} [\mathbf{H}_{W_T}(Z)], \mathbb{E}_{\Delta_T} [\Delta_T \Delta_T^T] \rangle \\ &= \frac{1}{2} \langle \mathbb{E}_{W_T, Z} [\mathbf{H}_{W_T}(Z)], \text{diag}(\sum_{t=1}^T \sigma_t^2) \rangle \quad (\text{A.20}) \\ &= \frac{\sum_{t=1}^T \sigma_t^2}{2} \text{Tr}(\mathbb{E}_{W_T, Z} [\mathbf{H}_{W_T}(Z)]), \end{aligned}$$

where Eq. (A.19) is by the zero mean of the perturbation, Eq. (A.20) is by the independence of the coordinates of  $\Delta_T$ ,  $\langle \cdot, \cdot \rangle$  denotes the inner product of two matrices,  $\text{diag}(A)$  is the diagonal matrix with element  $A$  and  $\text{Tr}(\cdot)$  is the trace of a matrix.

Under the condition  $\mathbb{E}_{W_T, S'} [\gamma(W_T, S')] \geq 0$ , we now bound  $\mathcal{E}_\mu(\mathcal{A})$  instead of its absolute

value,  $|\mathcal{E}_\mu(\mathcal{A})|$ . With the inequality  $\log(x+1) \leq x$ , the following is straightforward,

$$\begin{aligned} \mathcal{E}_\mu(\mathcal{A}) &\leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}} [\mathbb{V}(W_{t-1})] + \mathbb{E}_{W_T, S, S'} [\gamma(W_T, S) - \gamma(W_T, S')]} \\ &\leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}} [\mathbb{V}(W_{t-1})] + \mathbb{E}_{W_T, S} [\gamma(W_T, S)]} \\ &= \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}} [\mathbb{V}(W_{t-1})] + \frac{\sum_{t=1}^T \sigma_t^2}{2} \text{Tr}(\mathbb{E}_{W_T, Z} [\mathbf{H}_{W_T}(Z)])}. \end{aligned}$$

Since every choice of  $\sigma$  gives a valid generalization bound. The optimal bound can be found by simply utilizing the fact  $A/\sigma + \sigma^2 B \geq 3(A/2)^{2/3} B^{1/3}$  for any positive  $A$  and  $B$ , where the equality is achieved by the optimal  $\sigma$ . Finally, rearranging the terms will complete the proof.  $\square$

### A.3.5 Corollary A.3.2

**Corollary A.3.2.** *If the loss function is differentiable and  $\beta$ -smooth with respect to  $w$ , then,*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{R^2 d}{n} \sum_{t=1}^T \log \left( \frac{\lambda_t^2 \mathbb{E}_{W_{t-1}} [\mathbb{V}_t(W_{t-1})]}{d \sigma_t^2} + 1 \right)} + \beta d \sum_{t=1}^T \sigma_t^2.$$

*Proof.* Recall the smoothness implies  $f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|^2$  for all  $\mathbf{v}$  and  $\mathbf{w}$ . By the triangle inequality, we have

$$|\mathbb{E} [L_\mu(W_T) - L_\mu(W_T + \Delta_T)]| \leq |\mathbb{E} [\langle \nabla_w \ell(W_T, Z), \Delta_T \rangle]| + \frac{\beta}{2} \mathbb{E} [\|\Delta_T\|^2] = \frac{\beta d \sum_{t=1}^T \sigma_t^2}{2}$$

Thus, we can see that  $|\mathbb{E} [L_\mu(W_T) - L_\mu(W_T + \Delta_T)]| + |\mathbb{E} [L_S(W_T + \Delta_T) - L_S(W_T)]| \leq \beta d \sum_{t=1}^T \sigma_t^2$ .

This completes the proof.  $\square$



**Remark A.3.1.** In Corollary A.3.2, we note that the dependence of  $d$  in the bound results from the spherical Gaussian noise used in the construction of the weight process  $\widehat{W}_T$ . It is possible to replace the spherical Gaussian with a Gaussian noise having a non-diagonal covariance that reflects the geometry of the loss landscape. With this replacement, the dimension  $d$  in the flatness term will be replaced by the trace of  $\sum_{t=1}^T \kappa_t$ , where  $\kappa_t$  is the covariance matrix of the noise added at step  $t$ . Please refer to [153] for a similar development.

## A.4 Application in Neural Networks: Proof of Theorems 3.5.1 and 3.5.2

### A.4.1 Lemma A.4.1

**Lemma A.4.1.** Under the same conditions of Theorem 3.4.2, the generalization error of SGD is upper bounded by

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{3}{2} \left( \sum_{t=1}^T \frac{R^2 \lambda_t^2 T}{n} \mathbb{E} [\|g(W_{t-1}, B_t)\|_2^2] \mathbb{E} [\text{Tr}(\mathbf{H}_{W_T}(Z))] \right)^{\frac{1}{3}}$$

*Proof.* The proof of this lemma follows the same steps in the proof of Theorem 3.4.2 except that we require a different use of Lemma 3.4.2 here.

Specially, we let  $\Omega = 0$  in Lemma 3.4.2. The remaining steps are the same in the proof of Theorem 3.4.2 and should be straightforward.  $\square$

**Remark A.4.1.** The bound in Lemma A.4.1 is weaker than Eq. (3.4) in Theorem 3.4.2 as the centralized expected vector norm should be smaller than the original expected vector norm.

### A.4.2 Proof of Theorem 3.5.1

*Proof.* By Cauchy–Schwarz inequality, we have

$$\mathbb{E} [\|g(W_{t-1}, B_t)\|_2^2] = \mathbb{E} \left[ \left\| \frac{1}{b} \sum_{Z_i \in B_t} \nabla_w \ell(W_{t-1}, Z_i) \right\|_2^2 \right] \leq \mathbb{E} [\|\nabla_w \ell(W_{t-1}, Z)\|_2^2]. \quad (\text{A.21})$$

Notice that  $\nabla_w \ell(W, Z) = (W^{\mathbf{T}} X - Y)X$ . Let  $\hat{Y} = f(W, X)$ . Then,

$$\|\nabla_w \ell(W_{t-1}, Z)\|_2^2 = \|(W_{t-1}^{\mathbf{T}} X - Y)X\|_2^2 \leq \|W_{t-1}^{\mathbf{T}} X - Y\|_2^2 = (\hat{Y} - Y)^2 = 2\ell(W_{t-1}, Z). \quad (\text{A.22})$$

For the Hessian matrix, it's easy to see that  $\text{Tr}(\mathbf{H}_{W_T}(Z)) = 1$ .

Plugging everything into Lemma A.4.1, we have

$$\mathcal{E}_\mu(\mathcal{A}) \leq 3 \left( \sum_{t=1}^T \frac{R^2 \lambda_t^2 T}{4n} \mathbb{E}_{W_{t-1}, Z} [\ell(W_{t-1}, Z)] \right)^{\frac{1}{3}}.$$

This completes the proof.  $\square$

### A.4.3 Proof of Theorem 3.5.2

*Proof.* Since  $\nabla_{w_r} \ell(W, Z_i) = \frac{1}{\sqrt{m}} A_r(f(W, X_i) - Y_i) X_i \mathbb{I}_{r,i}$ , where  $\mathbb{I}_{r,i} = \mathbb{I}\{W_r^{\mathbf{T}} X_i \geq 0\}$ , we have

$$\|\nabla_w \ell(W_{t-1}, Z)\|_2^2 \quad (\text{A.23})$$

$$\begin{aligned} &= \sum_{r=1}^m \left\| \frac{1}{\sqrt{m}} A_r(f(W_{t-1}, X_i) - Y_i) X_i \mathbb{I}_{r,i} \right\|_2^2 \\ &\leq \sum_{r=1}^m \left\| \frac{1}{\sqrt{m}} A_r(f(W_{t-1}, X_i) - Y_i) \mathbb{I}_{r,i} \right\|_2^2 \\ &= \frac{1}{m} \sum_{r=1}^m \mathbb{I}_{r,i} \|f(W_{t-1}, X_i) - Y_i\|_2^2 \\ &= \frac{1}{m} \sum_{r=1}^m 2\mathbb{I}_{r,i} \ell(W_{t-1}, Z_i). \end{aligned} \quad (\text{A.24})$$

In addition, we notice that  $\text{Tr}(\mathbf{H}_{W_T}(Z)) = \frac{1}{m} \sum_{r=1}^m \mathbb{I}_{r,i,T}$ . Plugging everything into Lemma A.4.1, we have

$$\mathcal{E}_\mu(\mathcal{A}) \leq 3 \left( \sum_{t=1}^T \frac{R^2 \lambda_t^2 T}{4nm} \mathbb{E}_{W_{t-1}, Z_i} \left[ \sum_{r=1}^m \mathbb{I}_{r,i,t} \ell(W_{t-1}, Z_i) \right] \mathbb{E}_{W_T, Z_i} \left[ \frac{1}{m} \sum_{r=1}^m \mathbb{I}_{r,i,T} \right] \right)^{\frac{1}{3}}.$$

This completes the proof.  $\square$

## A.5 Experiment Details

### A.5.1 Architectures and Hyperparameters

In Section 3.6, MLP has one hidden layer with 512 hidden units, and AlexNet has five convolution layers (conv.  $3 \times 3$  (64 filters)  $\rightarrow$  max-pool  $3 \times 3 \rightarrow$  conv.  $5 \times 5$  (192 filters)  $\rightarrow$  max-pool  $3 \times 3 \rightarrow$  conv.  $3 \times 3$  (384 filters)  $\rightarrow$  conv.  $3 \times 3$  (256 filters)  $\rightarrow$  conv.  $3 \times 3$  (256 filters)  $\rightarrow$  max-pool  $3 \times 3$ ) followed by two fully connected layers both with 4096 units and a 10-way linear layer as the output layer. All of the convolution layers and the fully connected layers use standard rectified linear activation functions (ReLU).

The fixed learning rates used for MLP and AlexNet are 0.01 and 0.001, respectively. The batch size is set to 60. For the corrupted label experiment, we train the models until the models achieve 100% training accuracy. For other cases, we train the neural networks until the training loss converges (e.g.,  $< 0.0001$ ). Other settings are either described in Section 3.6 or apparent in the figures. Standard techniques such as weight decay and batch normalization are not used.

To choose the variance proxy  $R$ , we first collected all the per-instance losses  $\ell(W_{t-1}, Z_i)$  that were observed during training, then we let  $R = (\max_{i,t} \ell(W_{t-1}, Z_i) - \min_{i,t} \ell(W_{t-1}, Z_i))/2$  in our experiments.

In Section 3.7, we compare GMP with other advanced regularization methods. The results of other methods are reported directly from [244], and we now give their hyperparameter settings here for completeness. For Dropout, 10% of neurons are randomly selected to be deactivated in each layer. For label smoothing, the coefficient is 0.2. For flooding, the level is set to 0.02. For MixUp, we lineally combine random pairs of training data where the coefficient is drawn from Beta(1, 1). For adversarial training, the perturbation size is 1 for each pixel and we take one step to generate adversarial examples. For AMP, the number of inner iteration is 1, and the  $L_2$  norm ball radius values are 0.5 for PreActResNet18 and 0.1 for VGG16, respectively.

The implementation in this chapter is on PyTorch, and all the experiments are carried out on NVIDIA Tesla V100 GPUs (32 GB).

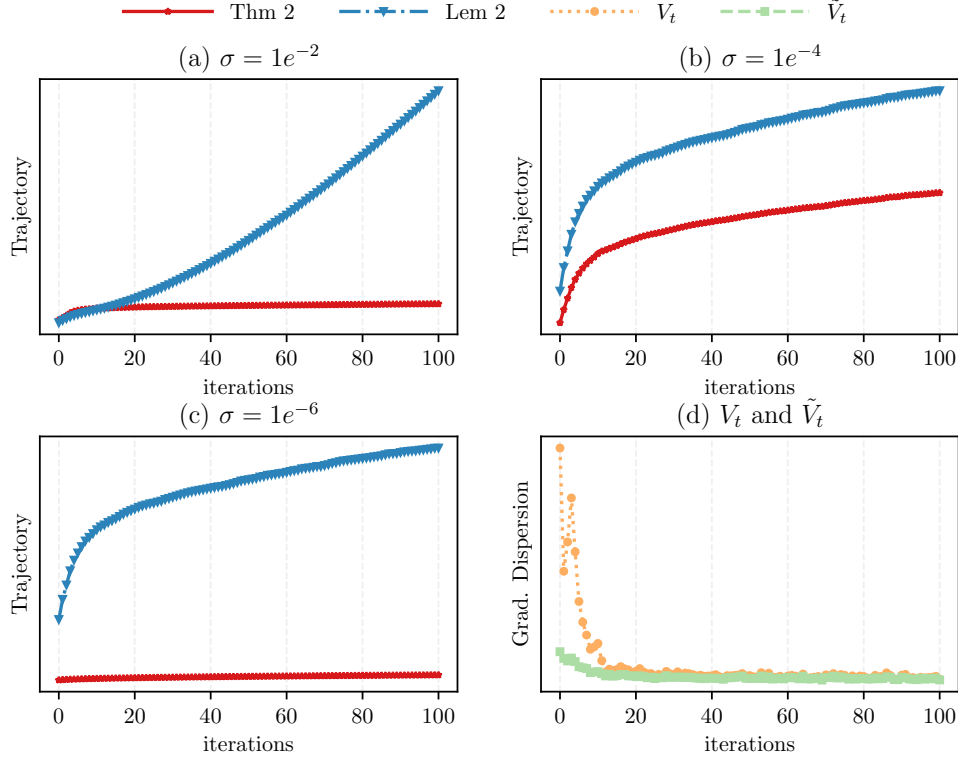


Figure A.1: Comparison Between Theorem 3.4.2 and Lemma 3.3.1 in stochastic setting.

### A.5.2 More Experiments on the Comparison Between Theorem 3.4.2 and Lemma 3.3.1

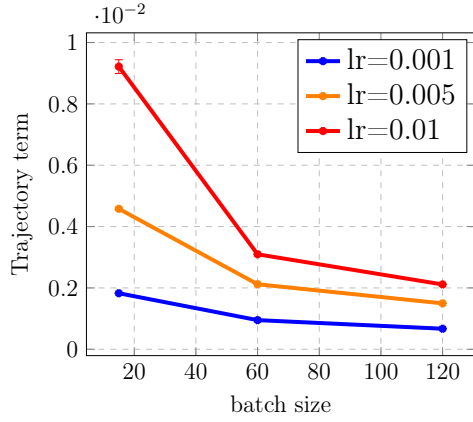
In Figure 3.1, we compare  $\sqrt{\frac{2}{n} \sum_{t=1}^T \mathbb{E} [\Psi(W_{t-1}) + \tilde{V}_t(W_{t-1})]}$  and  $\sqrt{\frac{1}{n} \sum_{t=1}^T \mathbb{E} [V_t(W_{t-1})]}$  for two different values of  $\sigma$ . Notice that we indeed use a weaker version of the trajectory term in Theorem 3.4.2, and the same constants like  $R$ ,  $\lambda_t$  and  $\sigma_t$  in the two bounds are ignored here. We choose the full dataset of MNIST and CIFAR10 to train the models with the fixed initialization. We also fix the sampling of batches, making the training completely deterministic. To estimate the local gradient sensitivity term, we randomly sample 20 Gaussian noises from  $\mathcal{N}(0, \sum_{\tau=1}^{t-1} \sigma_\tau \mathbf{I}_d)$  to perturb the model parameters and compute the average perturbed gradient. With this single draw and deterministic setting, the estimated  $\tilde{V}_t$  and  $V_t$  are the same, so the bound of Eq. (3.3) in Theorem 3.4.2 should be smaller than the bound in Lemma 3.3.1, as shown in Figure 3.1.

We also provide the experiments under the stochastic setting. Specially, We randomly choose 1/10 of the MNIST data and train the MLP model with a fixed learning rate and batch size. To estimate  $\mathbb{E}_{W_{t-1}, Z} [\nabla \ell(W_{t-1}, Z)]$ , we conduct 20 runs with different random seeds and save  $W$  after every iteration. In Figure A.1 (d), we can see that at the beginning of training, our gradient dispersion  $\mathbb{V}_t$  is much larger than  $\tilde{\mathbb{V}}_t$  in [153], but in the later training phase, the magnitude of gap between these two terms is very small. This is because the gradient norm will become tiny when  $W$  is near local minima. In Figure A.1 (a-c), we compare  $\sqrt{\frac{4}{n} \sum_{t=1}^T \frac{\lambda_t^2}{\sigma_t^2} \mathbb{E} [\Psi(W_{t-1}) + \tilde{\mathbb{V}}_t(W_{t-1})]}$  and  $\sqrt{\frac{d}{n} \sum_{t=1}^T \log \left( \frac{\lambda_t^2 \mathbb{E}[\mathbb{V}_t(W_{t-1})]}{d\sigma_t^2} + 1 \right)}$  under three different settings of  $\sigma$ . When  $\sigma$  is small (e.g.,  $\sigma = 1e^{-6}$ ), the local gradient sensitivity term will become small, but the factor  $1/\sigma^2$  will be very large, making the gap between  $\log(x+1)$  and  $x$  be extremely large, as shown in Figure A.1 (c). In this case, the improvement upon [153] is significant. When  $\sigma$  is large (e.g.,  $\sigma = 1e^{-2}$ ), at the beginning of training, the trajectory term in Lemma 3.3.1 will be smaller than the trajectory term in Theorem 3.4.2 since our  $\mathbb{V}_t$  is relatively large. However, since  $\Psi(W_{t-1})$  has the cumulative variance, and  $\mathbb{V}_t$  and  $\tilde{\mathbb{V}}_t$  become closer soon or later, the bound in Lemma 3.3.1 will be greater than the bound in Theorem 3.4.2 in the later training phase.

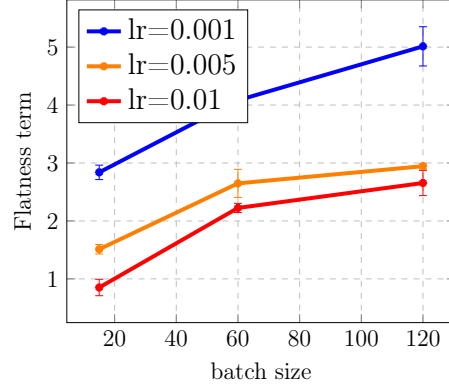
### A.5.3 Learning Rate and Batch Size.

The learning rate and batch size have some impact on Eq. (3.4) in Theorem 3.4.2. We now investigate this by performing experiments with varying learning rates and batch sizes. In our experiments, the model is continuously updated until the average training loss drops below 0.0001. We separate trajectory and flatness terms of the bound and plot them in Figure A.2.

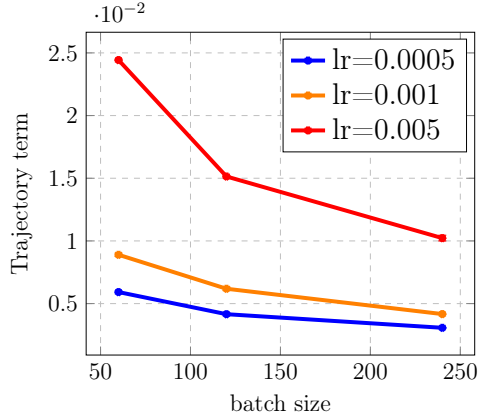
A key observation in Figure A.2 is that the learning rate impacts the trajectory term and the flatness term in opposite ways, as seen, for example, in (a) and (b), where the two set of curves swap their orders in the two figures. On the other hand, the batch size also impacts the two terms in opposite ways, as seen in (a) and (b) where curves decrease in (a) but increase in (b). This makes the generalization bound, i.e., the sum of the two terms, have a rather complex relationship with the settings of learning rate and batch size. This relationship is further complicated by the fact that a small learning rate requires a longer training time, or a larger number  $T$  of training iterations, which increases the number that are summed over in the trajectory term. Nonetheless, we do observe that a smaller batch size gives a lower value of the flatness term ((b) and (d)), confirming the previous wisdom that small batch sizes enable the neural network to find a flat minima [110].



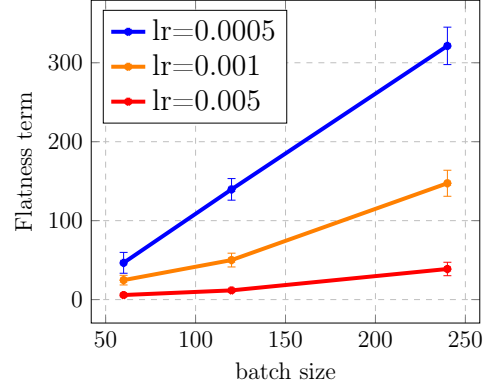
(a) MLP on MNIST



(b) MLP on MNIST



(c) AlexNet on CIFAR10



(d) AlexNet on CIFAR10

Figure A.2: The impact of learning rate and batch size on the trajectory term and the flatness term in Eq. (3.4).

#### A.5.4 Implementation of Theorem 3.5.1 and Theorem 3.5.2

We let  $d_0 = 200$  and use a two-layer ReLU network with hidden units 10000 to generate  $Y$ . Moreover, we apply a tanh function to the output of this network so that  $|Y| \leq 1$ . The input data  $X \sim \mathcal{N}(0, \mathbf{I})$  and we normalize  $X$  before training so that  $\|X\| = 1$ . For the training phase, we choose  $m = 100$ , let data size be 20,000, batch size be 100 and learning rate be 0.5. Figure A.3 and Figure A.4 compare the empirical generalization gap with the bound in Theorem 3.5.1 and that in Theorem 3.5.2, respectively.

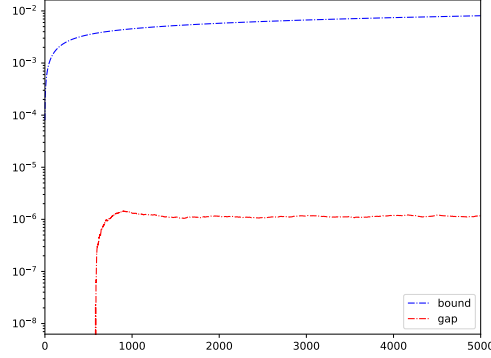


Figure A.3: Generalization gap of a linear network v.s. Theorem 3.5.1. Note y-axis is log-scale.

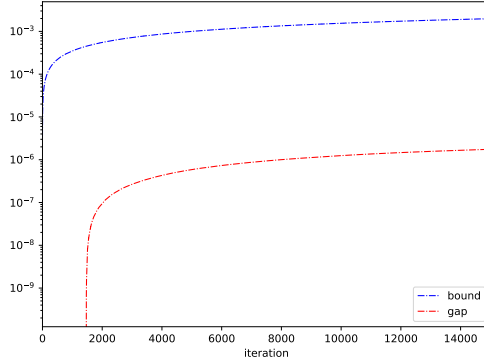


Figure A.4: Generalization gap of a two-layer ReLU network v.s. Theorem 3.5.2.

### A.5.5 Algorithm of Dynamic Gradient Clipping and Additional Results

The dynamic gradient clipping algorithm is described in Algorithm 1. For both MLP and AlexNet, we let  $\alpha = 0.1$ . The start step for clipping,  $T_c$ , is also an important hyperparameter. However, it can be removed by detecting the evolution of the average gradient norm for each epoch. Specifically, whenever the average gradient norm of epoch  $j$  is larger than the average gradient norm of epoch  $j - 1$ , the clipping operation begins.

From Figure 3.4 and Figure A.5 we can see that dynamic gradient clipping effectively alleviates overfitting by conspicuously slowing down the transition of training to the memorization regime, without changing the convergence speed of testing accuracy. Unfortunately, the current design of the dynamic gradient clipping algorithm does not provide a

---

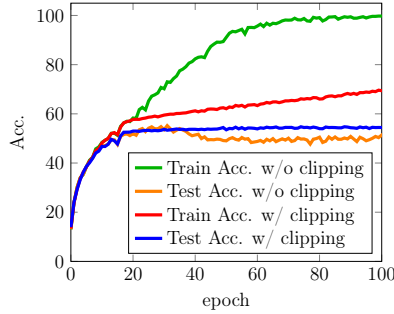
**Algorithm 1** Dynamic Gradient Clipping

---

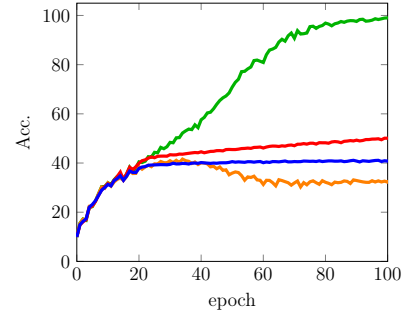
**Require:** Training set  $S$ , Batch size  $b$ , Loss function  $\ell$ , Initial model parameter  $\mathbf{w}_0$ , Learning rate  $\lambda$ , Initial minimum gradient norm  $\mathcal{G}$ , Number of iterations  $T$ , Clipping parameter  $\alpha$ , Clipping step  $T_c$

```
1: for  $t \leftarrow 1$  to  $T$  do
2:   Sample  $\mathcal{B} = \{\mathbf{z}_i\}_{i=1}^b$  from training set  $S$ 
3:   Compute gradient:
      $\mathbf{g}_{\mathcal{B}} \leftarrow \sum_{i=1}^b \nabla_{\mathbf{w}} \ell(\mathbf{w}_{t-1}, \mathbf{z}_i) / b$ 
4:   if  $t > T_c$  then
5:     if  $\|\mathbf{g}_{\mathcal{B}}\|_2 > \mathcal{G}$  then
6:        $\mathbf{g}_{\mathcal{B}} \leftarrow \alpha \cdot \mathcal{G} \cdot \mathbf{g}_{\mathcal{B}} / \|\mathbf{g}_{\mathcal{B}}\|_2$ 
7:     else
8:        $\mathcal{G} \leftarrow \|\mathbf{g}_{\mathcal{B}}\|_2$ 
9:     end if
10:  end if
11:  Update parameter:  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \lambda \cdot \mathbf{g}_{\mathcal{B}}$ 
12: end for
```

---



(a) noise=0.2 (CIFAR10)



(b) noise=0.4 (CIFAR10)

Figure A.5: Dynamic Gradient Clipping (AlexNet).

significant improvement for models trained on a clean dataset (without label noise). Designing better regularization algorithms may require understanding the epoch-wise double descent curve of gradient dispersion where the model is trained on a clean dataset.



### A.5.6 Discussion on Gradient Dispersion of Models Trained on Clean Datasets

In the case of no noise injected, Figure 3.3a shows that the model with good generalization property has a exponentially-decaying gradient dispersion. This is consistent with our discussion of Lemma 3.4.3 in Section 3.4, that is, small  $I(G_t + N_t; S | \widetilde{W}_{t-1})$  indicates good generalization. Notably, gradient dispersion of AlexNet trained on the real CIFAR10 data still has a epoch-wise double descent curve. The difference between Figure 3.3e with Figure 3.3f-3.3h is that the testing accuracy does not decrease in the second phase/memorization regime for AlexNet trained on the true CIFAR10 data. Loosely speaking, we conjugate that memorizing random labels will hurt the performance on unseen clean data but memorizing clean (or true) labels will not. This may explain why dynamic gradient clipping or preventing the training entering the memorization regime cannot improve the performance on a clean dataset.

### A.5.7 Algorithm of Gaussian Model Perturbation and Additional Results

---

#### Algorithm 2 Gaussian Model Perturbation Training

---

**Require:** Training set  $S$ , Batch size  $b$ , Loss function  $\ell$ , Initial model parameter  $\mathbf{w}_0$ , Learning rate  $\lambda$ , Number of noise  $k$ , Standard deviation of Gaussian distribution  $\sigma$ , Lagrange multiplier  $\rho$

**while**  $\mathbf{w}_t$  not converged **do**  
2:   Update iteration:  $t \leftarrow t + 1$   
   Sample  $\mathcal{B} = \{\mathbf{z}_i\}_{i=1}^b$  from training set  $S$   
4:   Sample  $\Delta_j \sim \mathcal{N}(0, \sigma^2)$  for  $j \in [k]$   
   Compute gradient:  
   
$$g_{\mathcal{B}} \leftarrow \sum_{i=1}^b \left( \nabla_{\mathbf{w}} \ell(\mathbf{w}_t, z_i) + \rho \sum_{j=1}^k (\nabla_{\mathbf{w}} \ell(\mathbf{w}_t + \Delta_j, z_i) - \nabla_{\mathbf{w}} \ell(\mathbf{w}_t, z_i)) / k \right) / b$$
  
6:   Update parameter:  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \lambda \cdot g_{\mathcal{B}}$   
**end while**

---

The GMP algorithm is given in Algorithm 2. Table 3.1 and Table A.1 show that our method is competitive to the state-of-the-art regularization techniques. Specifically, our method achieves the best performance on SVHN for both models and on CIFAR-100 where VGG16 is employed. Particularly, testing accuracy is improved by nearly 2% on the CIFAR-100 dataset with VGG16. For other tasks, GMP is always able to achieve the top-3

| Method                   | SVHN               | CIFAR-10           | CIFAR-100          |
|--------------------------|--------------------|--------------------|--------------------|
| ERM                      | 97.05±0.063        | 94.98±0.212        | 75.69±0.303        |
| Dropout                  | 97.20±0.065        | 95.14±0.148        | 75.52±0.351        |
| L.S.                     | 97.22±0.087        | 95.15±0.115        | 77.93±0.256        |
| Flooding                 | 97.16±0.047        | 95.03±0.082        | 75.50±0.234        |
| MixUp                    | 97.26±0.044        | 95.91±0.117        | 78.22±0.210        |
| Adv. Tr.                 | 97.23±0.080        | 95.01±0.085        | 74.77±0.229        |
| AMP                      | <b>97.70±0.025</b> | <b>96.03±0.091</b> | <b>78.49±0.308</b> |
| <b>GMP</b> <sup>3</sup>  | 97.43±0.037        | 95.64±0.053        | 78.05±0.208        |
| <b>GMP</b> <sup>10</sup> | 97.34±0.058        | 95.71±0.073        | 78.07±0.170        |

Table A.1: Top-1 classification accuracy acc.(%) of PreActResNet18. We run experiments 10 times and report the mean and the standard deviation of the testing accuracy. Super-script denotes the number of sampled Gaussian noises during training.

| Method                                 | SVHN        | CIFAR-10    | CIFAR-100   |
|----------------------------------------|-------------|-------------|-------------|
| <b>GMP</b> <sub>abs</sub> <sup>3</sup> | 97.10±0.054 | 94.21±0.139 | 74.80±0.113 |

Table A.2: Top-1 classification accuracy acc.(%) of VGG16.

performance. In addition, we find that increasing the number of sampled noises does not guarantee the improvement of testing accuracy and may even degrade the performance on some datasets (e.g., SVHN). This hints that we can use small number of noises to reduce the running time without losing performance. Moreover, we observe that **GMP with  $k = 3$  usually takes around  $1.76\times$  that of ERM training time**, which is affordable.

For PreActResNet18, the performance of GMP appears slightly inferior to the most recent record given by AMP [244]. Noting that the key ingredient of AMP, “max-pooling” in the parameter space, greatly resembles regularization term in GMP, which may be seen as “average-pooling” in the same space.

One potential extension of GMP is to let the variance of the noise distribution be a function of the iteration step  $t$ . In other words, using the time-dependent  $\sigma_t$  instead of a constant  $\sigma$ .

We also consider the following regularized scheme, which is a absolute value version.

$$\min_w L_s(w) + \rho \mathbb{E}_{\Delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)} \left[ \left| L_s(w + \Delta) - L_s(w) \right| \right].$$

This scheme can still perform well as shown in Table A.2. In fact, it outperforms GMP<sup>3</sup> on CIFAR-100. This hints that it is possible to improve the performance by choosing other norm of  $L_s(w + \Delta) - L_s(w)$ .

# Appendix B

## Appendix of Chapter 4

### B.1 Additional Background

#### B.1.1 Information-Theoretic Bounds for SGD

Recently, [153, 213] apply information-theoretic analysis to the generalization of models trained with SGD by invoking an auxiliary weight process (AWP). We now denote this auxiliary weight process by  $\mathcal{A}_{AWP}$ . Let  $\mathcal{A}_{SGD}$  be the original algorithm of SGD, [153, 213] obtain generalization bounds by the following construction,

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}_{SGD}) &= \mathcal{E}_\mu(\mathcal{A}_{SGD}) + \mathcal{E}_\mu(\mathcal{A}_{AWP}) - \mathcal{E}_\mu(\mathcal{A}_{AWP}) \\ &\leq \underbrace{\mathcal{O}\left(\sqrt{\frac{I(W_{AWP}; S)}{n}}\right)}_{\text{Lemma 2.4.11}} + \underbrace{|\mathcal{E}_\mu(\mathcal{A}_{SGD}) - \mathcal{E}_\mu(\mathcal{A}_{AWP})|}_{\text{residual term}},\end{aligned}\tag{B.1}$$

where  $W_{AWP}$  is the output hypothesis by  $\mathcal{A}_{AWP}$ .

Notably, it remains uncertain whether the residual term is sufficiently small for the information-theoretic bounds of  $\mathcal{A}_{AWP}$  to yield meaningful insights into SGD. Although there exists an optimal  $\mathcal{A}_{AWP}$  that tightens the bound in Eq. (B.1), finding such an optimal  $\mathcal{A}_{AWP}$  beyond the isotropic Gaussian noise covariance case is challenging. It's worth noting that [213] provides an optimal bound for the isotropic Gaussian noise case. Nevertheless, our empirical results, as illustrated in Figure 4.5, demonstrate that the bounds presented in this paper outperform the isotropic Gaussian noise case.

Table B.1: Comparison of the results in this work

|                                                                                                         | Bounds                                                                                                                                             | Remarks                                                                                          |
|---------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Trajectories-based Bounds. Pros: less assumptions, can track training dynamics; Cro: Time-Dependent     |                                                                                                                                                    |                                                                                                  |
| Theorem 4.3.1                                                                                           | $\mathcal{O}\left(\sqrt{\frac{d}{n}}\mathbb{E}\left[\log\frac{b_1}{d}-\frac{b_2}{d}\right]\right)$                                                 | Isotropic covariance for Gaussian prior                                                          |
| Corollary 4.3.1                                                                                         | $\mathcal{O}\left(\sqrt{\frac{d}{n}}\sum_{t=1}^T\mathbb{E}\log\left(\frac{\mathbb{E}\ G_t-\tilde{g}_t\ ^2}{d}+1\right)\right)$                     | Bound for langevin dynamic; tighter than [153, Prop. 3.]                                         |
| Theorem 4.3.2                                                                                           | $\mathcal{O}\left(\sqrt{\frac{1}{n}}\sum_{t=1}^T\mathbb{E}\left[\text{tr}\left\{\log\frac{\Sigma_t^*C_T^{-1}}{b}\right\}\right]\right)$            | Population GNC for prior; tighter than Thm. 4.3.1                                                |
| Terminal-State-based Bounds. Pro: time-indepdent; Cro: more assumptions, cannot track training dynamics |                                                                                                                                                    |                                                                                                  |
| Theorem 4.4.1                                                                                           | $\mathcal{O}\left(\sqrt{\frac{1}{n}}\mathbb{E}\left[\text{tr}\left\{\log\left(\Lambda_{W_S^*}^{-1}\Lambda_{w_\mu^*}\right)\right\}\right]\right)$  | General result; hard to measure in practice                                                      |
| Corollary 4.4.1                                                                                         | $\mathcal{O}\left(\sqrt{\frac{1}{n\eta}}\mathbb{E}\left[\text{tr}\left\{\log\left([H_{w^*}C_T^{-1}]\Lambda_{w_\mu^*}\right)\right\}\right]\right)$ | Under conditions: $H_{w^*}\Lambda_{w^*}=\Lambda_{w^*}H_{w^*}$ and $H_{w^*}\Sigma_T=\mathbf{I}_d$ |
| Corollary 4.4.2                                                                                         | $\mathcal{O}\left(\sqrt{\frac{d}{n}}\log\left(\frac{b}{\eta d}\mathbb{E}\ W_S^*-\hat{w}\ ^2+1\right)\right)$                                       | $\hat{w}$ is flexible; $\frac{2}{\eta}\gg\lambda_1$ ; other conditions same as Cor. 4.4.1        |
| Theorem 4.4.2                                                                                           | $\mathcal{O}\left(\mathbb{E}\sqrt{\frac{M^2b}{\eta}}\ W_S^*-W_{S_j}^*\ ^2\right)$                                                                  | Bounded loss; $\Lambda(W_{S_j}^*)=\Lambda(W_S^*)$ ; other conditions same as Cor. 4.4.2          |

In this paper, we do not attempt to find an optimal  $\mathcal{A}_{AWP}$ , but instead, we invoke the SDE approximation (i.e. Eq. (4.5)), denoted as  $\mathcal{A}_{SDE}$ . Formally,

$$\begin{aligned}
 \mathcal{E}_\mu(\mathcal{A}_{SGD}) &= \mathcal{E}_\mu(\mathcal{A}_{SGD}) + \mathcal{E}_\mu(\mathcal{A}_{SDE}) - \mathcal{E}_\mu(\mathcal{A}_{SDE}) \\
 &\leq \underbrace{\mathcal{O}\left(\sqrt{\frac{I(W_{SDE}; S)}{n}}\right)}_{\text{Lemma 2.4.11}} + \underbrace{|\mathcal{E}_\mu(\mathcal{A}_{SGD}) - \mathcal{E}_\mu(\mathcal{A}_{SDE})|}_{\text{residual term}}, \tag{B.2}
 \end{aligned}$$

where  $W_{SDE}$  is the output hypothesis by  $\mathcal{A}_{SDE}$ .

Empirical evidence from [223, 125] and our Figure 4.1 suggests that the residual term in Eq. (B.2) is small. This observation motivates our investigation into the generalization of SGD using the information-theoretic bounds of SDE directly.

## B.1.2 Summary of Results

The main generalization bounds obtained in this paper are summarized in Table B.1.

## B.1.3 Theoretical Validation of SDE

To theoretically assess the validation of SDE in approximating SGD, two essential technical definitions are necessary.

**Definition B.1.1** (Test Functions). *Class  $\mathcal{F}$  of continuous functions  $\mathbb{R}^d \rightarrow \mathbb{R}$  has polynomial growth if  $\forall f \in \mathcal{F}$ , if there exists constants  $K, \kappa > 0$  s.t.  $|f(x)| < K(1 + |x|^\kappa)$  for all  $x \in \mathbb{R}$ .*

**Definition B.1.2** (Order- $\alpha$  weak approximation). *Let  $\eta \in (0, 1)$ ,  $T > 0$  and  $N = \lfloor T/\eta \rfloor$ . Let  $\mathcal{F}$  be the set of test Functions. We say that the SDE in Eq. (4.4) is an order  $\alpha$  weak approximation of the SGD in Eq. (4.1) if for every  $f \in \mathcal{F}$ , there exists  $C > 0$ , independent of  $\eta$ , s.t. for all  $k = 0, 1, \dots, N$ ,*

$$|\mathbb{E}[f(\omega_{k\eta})] - \mathbb{E}[f(W_k)]| \leq C\eta^\alpha.$$

Below is a classical result.

**Lemma B.1.1** ([122, Theorem 1]). *Let  $\eta \in (0, 1)$ ,  $T > 0$  and  $N = \lfloor T/\eta \rfloor$ . Let  $\mathcal{F}$  be the set of functions of polynomial growth, i.e.  $f \in \mathcal{F}$ , if there exists constants  $K, \kappa > 0$  s.t.  $|f(x)| < K(1 + |x|^\kappa)$ . Assume  $\nabla \ell$  is Lipschitz continuous, has at most linear asymptotic growth and has sufficiently high derivatives belonging to  $\mathcal{F}$ , then SDE in Eq. (4.4) is an order 1 weak approximation of the SGD in Eq. (4.1). Or equivalently, for every  $f \in \mathcal{F}$ , there exists  $C > 0$ , independent of  $\eta$ , s.t. for all  $k = 0, 1, \dots, N$ ,  $|\mathbb{E}[f(\omega_{k\eta})] - \mathbb{E}[f(W_k)]| < C\eta$ .*

This theorem suggests that SGD and SDE closely track each other when they result in similar distributions of outcomes, such as the returned hypothesis  $W$ . In addition, the closeness of distributions is formulated through expectations of suitable classes of test functions, as defined in Definition B.1.1. As mentioned in [125], of particular interest for machine learning are test functions like generalization error  $\mathcal{E}_\mu$ , which may not adhere to formal conditions such as differentiability assumed in classical theory but are still valuable for experimental use. Other typical choices of test functions includes weight norm, gradient norm, and the trace of noise covariance.

## B.1.4 Gaussian Distribution around Local Minimum

A multi-dimensional Ornstein-Uhlenbeck process is defined as

$$dx_t = -\mathbf{H}x_t dt + \mathbf{B}d\theta_t, \tag{B.3}$$

where  $x_t \in \mathbb{R}^d$ ,  $\mathbf{H}$ ,  $\mathbf{B}$  are  $d \times d$  matrices and  $\theta_t$  is an  $d$ -dimensional Wiener process.

Denote the density function of  $x_t$  as  $P(x, t)$ , then the corresponding Fokker-Planck equation describes the evolution of  $P(x, t)$ :

$$\frac{\partial P(x, t)}{\partial t} = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial}{\partial x_i} \left( P(x, t) \sum_{j=1}^d \mathbf{H}_{i,j} x_j \right) + \sum_{i=1}^d \sum_{j=1}^d \mathbf{D}_{i,j} \frac{\partial^2 P(x, t)}{\partial x_i \partial x_j},$$

where  $\mathbf{D} = \mathbf{B}\mathbf{B}^T/2$ .

Moreover, if  $\mathbf{H}$  is positive definite, then a stationary solution of  $P$  is given by [59]:

$$P(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left( -\frac{1}{2} x^T \Sigma^{-1} x \right), \quad (\text{B.4})$$

where  $\Sigma = \mathbb{E} [xx^T]$  is the covariance matrix of  $x$ .

When  $w$  is close to any local minimum  $w^*$ , we can use a second-order Taylor expansion to approximate the value of the loss at  $w$ ,

$$L_s(w) \approx L_s(w^*) + \frac{1}{2} (w - w^*)^T H_{w^*} (w - w^*). \quad (\text{B.5})$$

In this case, when  $w_t \rightarrow w^*$ , we have  $G_t = \nabla L_s(w_t) = H_{w^*} (w_t - w^*)$ . Recall Eq. (4.2), then

$$w_t = w_{t-1} - \eta G_t + \eta V_t = w_{t-1} - \eta H_{w^*} (w_{t-1} - w^*) + \eta V_t.$$

Let  $W'_t \triangleq W_t - w^*$  and recall Eq (B.3), we thus have the Ornstein-Uhlenbeck process for  $x_t = W'_t$  as

$$dW'_t = -\eta H_{w^*} W'_t dt + \eta \sqrt{C_t} d\theta_t. \quad (\text{B.6})$$

By Eq. (B.4), we have

$$P(W') \propto \exp \left( -\frac{1}{2} W'^T \Lambda_{w^*}^{-1} W' \right).$$

Consequently, the stationary distribution of  $W$  for a given  $w^*$  is  $\mathcal{N}(w^*, \Lambda_{w^*})$ .

For discrete case, we have

$$\begin{aligned} w'_t &= (\mathbf{I}_d - \eta H_{w^*}) w'_{t-1} + \eta V_t \\ &= (\mathbf{I}_d - \eta H_{w^*})^2 w'_{t-2} + \eta ((\mathbf{I}_d - \eta H_{w^*}) V_{t-1} + V_t) \\ &\vdots \\ &= \bar{H}^t w'_0 + \eta \sum_{i=0}^t \bar{H}^i V_{t-i}, \end{aligned}$$

where  $\bar{H} = \mathbf{I}_d - \eta H_{w^*}$ . Notably, when  $t$  is sufficiently large, then the first term is negligible, especially with a small learning rate, we have  $w'_t = w_t - w^* = \eta \sum_{i=0}^t \bar{H}^i V_{t-i}$ . When  $C_t$  does not change in the long time limit, then  $W'_t$  is the weighted sum of independent Gaussian random variables, which follows a Gaussian distribution, namely  $w_t \sim \mathcal{N}(w^*, \Lambda_{w^*})$ . We refer readers to [128, Theorem 1-2.] for a relaxed analysis in the discrete case.

## B.2 Omitted Proofs and Additional Results in Section 4.3

### B.2.1 Useful Lemmas

We first unroll the terminal parameters' mutual information  $I(W_T; S)$  to the full trajectories' mutual information via the lemma below.

**Lemma B.2.1.**  $I(W_T; S) \leq \sum_{t=1}^T I(-G_t + C_t^{1/2} N_t; S | W_{t-1})$ .

This lemma can be proved by recurrently applying the data processing inequality (DPI) and chain rule of the mutual information [170].

*Proof.* Recall the SDE approximation of SGD, i.e., Eq (4.5), we then have,

$$\begin{aligned} I(W_T; S) &= I(W_{T-1} - \eta G_T + \eta C_T^{1/2} N_T; S) \\ &\leq I(W_{T-1}, -\eta G_T + \eta C_T^{1/2} N_T; S) \end{aligned} \tag{B.7}$$

$$= I(W_{T-1}; S) + I(-\eta G_T + \eta C_T^{1/2} N_T; S | W_{T-1}) \tag{B.8}$$

$\vdots$

$$\leq \sum_{t=1}^T I(-\eta G_t + \eta C_t^{1/2} N_t; S | W_{t-1})$$

$$= \sum_{t=1}^T I(-G_t + C_t^{1/2} N_t; S | W_{t-1}).$$

where Eq. (B.7) is by the data processing inequality (e.g.,  $Z - (X, Y) - (X + Y)$  form a markov chain then  $I(X + Y, Z) \leq I(X, Y; Z)$ ), Eq. (B.8) is by the chain rule of the mutual information, and learning rate  $\eta$  is dropped since mutual information is scale-invariant.  $\square$



The following lemma is inspired by the classic Log sum inequality in [41, Theorem 2.7.1].

**Lemma B.2.2.** *For non-negative numbers  $\{a_i\}_{i=1}^n$  and  $\{b_i\}_{i=1}^n$ ,*

$$\sum_{i=1}^n b_i \log \frac{a_i}{b_i} \leq \left( \sum_{i=1}^n b_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i},$$

*with equality if and only if  $\frac{a_i}{b_i} = \text{const.}$*

*Proof.* Since  $\log$  is a concave function, according to Jensen's inequality, we have

$$\sum_{i=1}^n \alpha_i \log(x_i) \leq \log\left(\sum_{i=1}^n \alpha_i x_i\right),$$

where  $\sum_{i=1}^n \alpha_i = 1$ .

Let  $\alpha_i = \frac{b_i}{\sum_{i=1}^n b_i}$  and  $x_i = \frac{a_i}{b_i}$ , and plugging them into the inequality above, we have

$$\sum_{i=1}^n \frac{b_i}{\sum_{i=1}^n b_i} \log\left(\frac{a_i}{b_i}\right) \leq \log\left(\sum_{i=1}^n \frac{b_i}{\sum_{i=1}^n b_i} \frac{a_i}{b_i}\right) = \log\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right),$$

which implies

$$\sum_{i=1}^n b_i \log\left(\frac{a_i}{b_i}\right) \leq \left(\sum_{i=1}^n b_i\right) \log\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right).$$

This completes the proof. □

## B.2.2 Proof of Lemma 4.3.1

*Proof.* For any  $t \in [T]$ , similar to the proof of Lemma 2.4.10 in [170]:

$$\begin{aligned} & I(-G_t + C_t^{1/2} N_t; S | W_{t-1} = w_{t-1}) \\ &= \mathbb{E}_S^{w_{t-1}} \left[ \text{D}_{\text{KL}}(Q_{\hat{G}_t | w_{t-1}, S} || Q_{\hat{G}_t | w_{t-1}}) \right] \\ &= \mathbb{E}_S^{w_{t-1}} \left[ \text{D}_{\text{KL}}(Q_{\hat{G}_t | w_{t-1}, S} || P_{\hat{G}_t | w_{t-1}}) - \text{D}_{\text{KL}}(Q_{\hat{G}_t | w_{t-1}} || P_{\hat{G}_t | w_{t-1}}) \right] \\ &\leq \mathbb{E}_S^{w_{t-1}} \left[ \text{D}_{\text{KL}}(Q_{\hat{G}_t | w_{t-1}, S} || P_{\hat{G}_t | w_{t-1}}) \right], \end{aligned} \tag{B.9}$$

where Eq. (B.9) is due to the fact that KL divergence is non-negative, and the equality holds when  $P_{\hat{G}_t|w_{t-1}} = Q_{\hat{G}_t|w_{t-1}}$  for  $W_{t-1} = w_{t-1}$ .

Thus, we conclude that

$$I(\hat{G}_t; S|W_{t-1} = w_{t-1}) = \inf_{P_{\hat{G}_t|w_{t-1}}} \mathbb{E}_S^{w_{t-1}} \left[ D_{\text{KL}}(Q_{\hat{G}_t|w_{t-1}, S} || P_{\hat{G}_t|w_{t-1}}) \right].$$

Taking expectation over  $W_{t-1}$  for both side above, we have

$$I(\hat{G}_t; S|W_{t-1}) = \mathbb{E}_{W_{t-1}} \left[ \inf_{P_{\hat{G}_t|W_{t-1}}} \mathbb{E}_S^{W_{t-1}} \left[ D_{\text{KL}}(Q_{\hat{G}_t|W_{t-1}, S} || P_{\hat{G}_t|W_{t-1}}) \right] \right].$$

This completes the proof.  $\square$

### B.2.3 Proof of Theorem 4.3.1

*Proof.* We first prove Eq. (4.6). Recall Lemma 4.3.1 and assume  $C_t$  is a positive-definite matrix, for any  $t \in [T]$ , we have

$$\begin{aligned} & I(-G_t + C_t^{1/2} N_t; S|W_{t-1} = w_{t-1}) \\ & \leq \inf_{\tilde{g}_t, \sigma_t} \mathbb{E}_S^{w_{t-1}} \left[ D_{\text{KL}}(Q_{-G_t + C_t^{1/2} N_t|w_{t-1}, S} || P_{-\tilde{g}_t + \sigma_t N_t|w_{t-1}}) \right] \\ & = \inf_{\tilde{g}_t, \sigma_t} \mathbb{E}_S^{w_{t-1}} \left[ \frac{1}{2} \left[ \log \frac{\det(\sigma_t^2 \mathbf{I}_d)}{\det(C_t)} - d + \frac{1}{\sigma_t^2} ((G_t - \tilde{g}_t)^T \mathbf{I}_d^{-1} (G_t - \tilde{g}_t)) + \frac{1}{\sigma_t^2} \text{tr} \{ \mathbf{I}_d^{-1} C_t \} \right] \right] \end{aligned} \quad (\text{B.10})$$

$$= \frac{1}{2} \inf_{\tilde{g}_t, \sigma_t} \mathbb{E}_S^{w_{t-1}} \left[ \frac{1}{\sigma_t^2} (\|G_t - \tilde{g}_t\|^2 + \text{tr} \{C_t\}) + d \log \sigma_t^2 - d - \text{tr} \{ \log C_t \} \right], \quad (\text{B.11})$$

where Eq. (B.10) is by Eq. (2.1), Eq. (B.11) is due to the fact that  $\log \det(C_t) = \text{tr} \{ \log C_t \}$  when  $C_t$  is positive definite.

Recall that  $h_1(w) = \mathbb{E}_S^w [\|G_t - \tilde{g}_t\|^2 + \text{tr} \{C_t\}]$  and  $h_2(w) = \mathbb{E}_S^w [\text{tr} \{ \log C_t \}]$ , then we have

$$\begin{aligned} & \frac{1}{2} \inf_{\tilde{g}_t, \sigma_t} \frac{1}{\sigma_t^2} \mathbb{E}_S^{w_{t-1}} [\|G_t - \tilde{g}_t\|^2 + \text{tr} \{C_t\}] + d \log \sigma_t^2 - d - \mathbb{E}_S^{w_{t-1}} [\text{tr} \{ \log C_t \}] \\ & \leq \frac{1}{2} \inf_{\sigma_t > 0} \frac{1}{\sigma_t^2} h_1(w_{t-1}) + d \log \sigma_t^2 - d - h_2(w_{t-1}) \\ & = \frac{1}{2} d \log \frac{h_1(w_{t-1})}{d} - \frac{1}{2} h_2(w_{t-1}), \end{aligned}$$

where we fix an arbitrary  $\tilde{g}_t$  and use the optimal  $\sigma^* = \sqrt{\frac{h_1(w_{t-1})}{d}}$ .

Plugging everything into Lemma B.2.1 and Lemma 2.4.11 will obtain Eq. (4.6).

We then prove the second part. Let  $\tilde{g}_t = \mathbb{E}_Z [\nabla \ell(w_{t-1}, Z)]$ , then

$$\begin{aligned} h_1(W_{t-1}) &= \mathbb{E}_S^{W_{t-1}} [\|G_t - \tilde{g}_t\|^2 + \text{tr} \{C_t\}] \\ &= \mathbb{E}_S^{W_{t-1}} [\text{tr} \{(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^{\mathbf{T}})\}] + \text{tr} \{\mathbb{E}_S^{W_{t-1}} [C_t]\} \\ &= \frac{1}{n} \text{tr} \{\Sigma_t^\mu\} + \frac{n-b}{b(n-1)} \text{tr} \{\mathbb{E}_S^{W_{t-1}} [\Sigma_t]\} \end{aligned} \quad (\text{B.12})$$

$$\begin{aligned} &= \frac{1}{n} \text{tr} \{\Sigma_t^\mu\} + \frac{n-b}{bn} \text{tr} \{\Sigma_t^\mu\} \\ &= \frac{1}{b} \text{tr} \{\Sigma_t^\mu\}, \end{aligned} \quad (\text{B.13})$$

where Eq. (B.12) is by  $\mathbb{E}_S [(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^{\mathbf{T}})] = \frac{1}{n} \Sigma_t^\mu$  for a given  $W_{t-1} = w_{t-1}$  and  $C_t = \frac{n-b}{b(n-1)} \Sigma_t$ , and Eq. (B.13) is by  $\mathbb{E}_S [\Sigma_t] = \frac{n-1}{n} \Sigma_t^\mu$ . This completes the proof.  $\square$

**Remark B.2.1.** *Unlike information-theoretic generalization bounds of SGLD in the literature, learning rate does not explicitly appear in the bound of Theorem 4.3.1. This is because the noise random variable has a tunable scaling factor in SGLD while the noise random variable has a fixed scaling factor  $\eta$ . The latter scaling factor is then dropped since mutual information is scale-invariant (see the proof of Lemma B.2.1 for more details).*

## B.2.4 Proof of Corollary 4.3.1

*Proof.* Let  $C_t = \mathbf{I}_d$ , by Theorem 4.3.1,

$$\begin{aligned} \mathcal{E}_\mu(\mathcal{A}) &\leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T d \mathbb{E}_{W_{t-1}} \left[ \log \frac{\mathbb{E}_S^{W_{t-1}} [\|G_t - \tilde{g}_t\|^2 + \text{tr} \{C_t\}]}{d} \right] - \mathbb{E}_{W_{t-1}, S} [\text{tr} \{\log C_t\}]} \\ &= \sqrt{\frac{R^2}{n} \sum_{t=1}^T d \mathbb{E}_{W_{t-1}} \left[ \log \frac{\mathbb{E}_S^{W_{t-1}} [\|G_t - \tilde{g}_t\|^2 + d]}{d} \right]} \\ &= \sqrt{\frac{R^2}{n} \sum_{t=1}^T d \mathbb{E}_{W_{t-1}} \left[ \log \frac{\mathbb{E}_S^{W_{t-1}} [\|G_t - \tilde{g}_t\|^2]}{d} + 1 \right]}. \end{aligned}$$

This completes the proof.  $\square$

### B.2.5 Proof of Theorem 4.3.2

*Proof.* Recall Lemma 4.3.1, we have

$$\begin{aligned}
& I(-G_t + C_t^{1/2} N_t; S | W_{t-1} = w_{t-1}) \\
& \leq \inf_{\tilde{c}_t} \mathbb{E}_S^{w_{t-1}} \left[ D_{\text{KL}}(Q_{\hat{G}_t | w_{t-1}, S} || P_{\hat{G}_t | w_{t-1}}) \right] \\
& = \inf_{\tilde{c}_t} \mathbb{E}_S^{w_{t-1}} \left[ \frac{1}{2} \left[ \log \frac{\det(\tilde{c}_t \Sigma_t^\mu)}{\det(C_t)} - d + \frac{1}{\tilde{c}_t} ((G_t - \tilde{g}_t)^\mathbf{T} (\Sigma_t^\mu)^{-1} (G_t - \tilde{g}_t)) + \frac{1}{\tilde{c}_t} \text{tr} \left\{ (\Sigma_t^\mu)^{-1} C_t \right\} \right] \right] \\
& = \frac{1}{2} \inf_{\tilde{c}_t} \frac{1}{\tilde{c}_t} \text{tr} \left\{ (\Sigma_t^\mu)^{-1} \mathbb{E}_S^{w_{t-1}} [(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^\mathbf{T})] \right\} \\
& \quad + \frac{1}{\tilde{c}_t} \text{tr} \left\{ (\Sigma_t^\mu)^{-1} \mathbb{E}_S^{w_{t-1}} [C_t] \right\} + \text{tr} \{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \} + d \log \tilde{c}_t - d \\
& = \frac{1}{2} \inf_{\tilde{c}_t} \frac{1}{\tilde{c}_t n} \text{tr} \left\{ (\Sigma_t^\mu)^{-1} \Sigma_t^\mu \right\} + \frac{n-b}{\tilde{c}_t b n} \text{tr} \left\{ (\Sigma_t^\mu)^{-1} \Sigma_t^\mu \right\} + \text{tr} \{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \} + d \log \tilde{c}_t - d \\
& \tag{B.14} \\
& = \frac{1}{2} \inf_{\tilde{c}_t} \frac{d}{\tilde{c}_t n} + \frac{(n-b)d}{\tilde{c}_t b n} + \text{tr} \{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \} + d \log \tilde{c}_t - d \\
& = \frac{1}{2} \inf_{\tilde{c}_t} \frac{d}{b \tilde{c}_t} + d \log \tilde{c}_t + \text{tr} \{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \} - d \\
& = \frac{d}{2} \log \frac{1}{b} + \frac{1}{2} \text{tr} \{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \},
\end{aligned}$$

where the last equality hold when  $\tilde{c}_t^* = 1/b$  and Eq. (B.14) is by

$$\mathbb{E}_S^{w_{t-1}} [(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^\mathbf{T})] = \frac{1}{n} \Sigma_t^\mu, \quad \text{and}$$

$$\mathbb{E}_S^{w_{t-1}} [C_t] = \frac{n-b}{b(n-1)} \mathbb{E}_S^{w_{t-1}} [\Sigma_t] = \frac{n-b}{b(n-1)} \frac{n-1}{n} \Sigma_t^\mu = \frac{n-b}{bn} \Sigma_t^\mu.$$

This completes the proof.  $\square$

### B.2.6 Proof of Lemma 4.3.2

*Proof.* Let the diagonal element of  $\Sigma_t^\mu/b$  in dimension  $k$  be  $a_k$ , then

$$\sum_{k=1}^d \log a_k \leq \left( \sum_{k=1}^d 1 \right) \cdot \log \left( \sum_{k=1}^d a_k \right) / \left( \sum_{k=1}^d 1 \right) = d \log(\text{tr} \{ \Sigma_t^\mu \} / bd),$$

where we invoke Lemma B.2.2.

This completes the proof.  $\square$

**Comparison with the work of [208]** [208] studies the algorithm of SGD with anisotropic noise, while our SDE analysis focuses on GD with anisotropic noise. This means that the discrete gradient noise arising from mini-batch sampling still exists in their analyzed algorithm, whereas the gradient noise is fully modeled as Gaussian in our Section B.1.1. Moreover, [208] uses matrix analysis tools to optimize the prior distribution. A significant distinction lies in their optimization analysis, which relies on the assumption that the trace of gradient noise covariance remains unchanged during training (see **Constrant 1** in their paper). Additionally, their final optimal posterior covariance is derived based on the assumption that the posterior distribution of  $W$  is invariant to the data index, see Assumption 1 in their paper. In contrast, our Section B.1.1 avoids making these assumptions and demonstrates the superiority of population gradient noise covariance (GNC) in Lemma 4.3.2, by invoking a variant of the log-sum inequality. In summary, our proof is simpler and more straightforward, while [208] makes a stronger claim about the optimality of population GNC based on their additional assumptions.

## B.3 Omitted Proofs and Additional Results in Section 4.4

In fact, this section provides a PAC-Bayes type analysis. The connection between information-theoretic bounds and PAC-Bays bounds have already been discussed in many previous works [17, 89, 6]. Roughly speaking, the most significant component of a PAC-Bayes bound is the KL divergence between the posterior distribution of a randomized algorithm output and a prior distribution, i.e.  $D_{\text{KL}}(Q_{W_T|S}||P_N)$  for some prior  $P_N$ . In essence, information-theoretic bounds can be view as having the same spirit. For concreteness, in Lemma 2.4.11,  $I(W_T; S) = \mathbb{E}_S[D_{\text{KL}}(Q_{W_T|S}||P_{W_T})]$ , in which case the marginal  $P_{W_T}$  is used as a prior of the algorithm output. Furthermore, by using Lemma 2.4.10, we have  $I(W_T; S) \leq \inf_{P_N} \mathbb{E}_S[D_{\text{KL}}(Q_{W_T|S}||P_N)]$ . Hence, Lemma 2.4.11 can be regarded as a PAC-Bayes bound with the optimal prior. In addition, the PAC-Bayes framework is usually used to provide a high-probability bound, while information-theoretic analysis is applied to bounding the expected generalization error. In this sense, information-theoretic framework is closer to another concept called MAC-Bayes [68].

### B.3.1 Proof of Lemma 4.4.1

*Proof.* Recall  $G_t = \nabla L_s(w_t) = H_{w^*}(w_t - w^*)$ . and Eq. (4.2), then

$$\begin{aligned} w_t &= w_{t-1} - \eta G_t + \eta V_t \\ &= w_{t-1} - \eta H_{w^*}(w_{t-1} - w^*) + \eta V_t. \end{aligned}$$

Let  $W'_t \triangleq W_t - w^*$ . Thus, as  $T \rightarrow \infty$ ,

$$\begin{aligned} &\mathbb{E}_{W'_T} [W'_T W'^{\mathbf{T}}_T] \\ &= \mathbb{E}_{W'_{T-1}, V_T} [(W'_{T-1} - \eta H_{w^*} W'_{T-1} + \eta V_t) (W'_{T-1} - \eta H_{w^*} W'_{T-1} + \eta V_t)^{\mathbf{T}}] \\ &= \mathbb{E}_{W'_{T-1}} [W'_{T-1} W'^{\mathbf{T}}_{T-1} - \eta H_{w^*} W'_{T-1} W'^{\mathbf{T}}_{T-1} - \eta W'_{T-1} W'^{\mathbf{T}}_{T-1} H_{w^*} + \eta^2 H_{w^*} W'_{T-1} W'^{\mathbf{T}}_{T-1} H_{w^*}] \\ &\quad + \eta^2 \mathbb{E}_{V_T} [V_T V_T^{\mathbf{T}}], \end{aligned}$$

where the last equation is by  $\mathbb{E}_{V_T}^{w_{T-1}} [V_T] = 0$ .

Recall that  $\mathbb{E}_{V_T} [V_T V_T^{\mathbf{T}}] = C_T$  and notice that  $\mathbb{E}_{W'_T} [W'_T W'^{\mathbf{T}}_T] = \mathbb{E}_{W'_{T-1}} [W'_{T-1} W'^{\mathbf{T}}_{T-1}] = \Lambda_{w^*}$  when  $T \rightarrow \infty$  (i.e. ergodicity), we have

$$\Lambda_{w^*} H_{w^*} + H_{w^*} \Lambda_{w^*} - \eta H_{w^*} \Lambda_{w^*} H_{w^*} = \eta C_T.$$

Furthermore, if  $H_{w^*}$  and  $\Lambda_{w^*}$  commute, namely  $\Lambda_{w^*} H_{w^*} = H_{w^*} \Lambda_{w^*}$ , we have

$$[H_{w^*} (2\mathbf{I}_d - \eta H_{w^*})] \Lambda_{w^*} = \eta C_T,$$

which will give use  $\Lambda_{w^*} = \eta [H_{w^*} (2\mathbf{I}_d - \eta H_{w^*})]^{-1} C_T$ .

This completes the proof.  $\square$

### B.3.2 Theorem B.3.1: A General Bound

The following bound can be easily proved by using Eq. (2.1).

**Theorem B.3.1.** *Under the same conditions in Lemma 2.4.11 and Lemma 4.4.1, then for any  $P_{W_T} = \mathcal{N}(\tilde{w}, \tilde{\Lambda})$ , where  $\tilde{w}$  and  $\tilde{\Lambda}$  are independent of  $S$ , we have*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2}{2n} \inf_{\tilde{w}, \tilde{\Lambda}} \mathbb{E}_{S, W_S^*} \left[ \log \frac{\det(\tilde{\Lambda})}{\det(\Lambda_{W_S^*})} + \text{tr} \left\{ \tilde{\Lambda}^{-1} \Lambda_{W_S^*} - \mathbf{I}_d \right\} + d_M^2(W_S^*, \tilde{w}; \tilde{\Lambda}) \right]},$$

where  $d_M(x, y; \Sigma) \triangleq \sqrt{(x - y)^{\mathbf{T}} \Sigma^{-1} (x - y)}$  is the Mahalanobis distance.

### B.3.3 Proof of Theorem 4.4.1

*Proof.* Let  $P_{W_T} = \mathcal{N}(w_\mu^*, \Lambda_{w_\mu^*})$ , then

$$\begin{aligned}
& \mathbb{E}_{S, W_S^*} \left[ \log \frac{\det(\Lambda_{w_\mu^*})}{\det(\Lambda_{W_S^*})} + \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} \Lambda_{W_S^*} - \mathbf{I}_d \right\} + (W_S^* - w_\mu^*)^\mathbf{T} \Lambda_{w_\mu^*}^{-1} (W_S^* - w_\mu^*) \right] \\
&= \mathbb{E}_{S, W_S^*} \left[ \log \frac{\det(\Lambda_{w_\mu^*})}{\det(\Lambda_{W_S^*})} + \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} \Lambda_{W_S^*} - \mathbf{I}_d \right\} + \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} (W_S^* - w_\mu^*) (W_S^* - w_\mu^*)^\mathbf{T} \right\} \right] \\
&= \mathbb{E}_{S, W_S^*} \left[ \log \frac{\det(\Lambda_{w_\mu^*})}{\det(\Lambda_{W_S^*})} \right] + \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} \mathbb{E}_{S, W_S^*} [\Lambda_{W_S^*}] - \mathbf{I}_d + \Lambda_{w_\mu^*}^{-1} \mathbb{E}_{W_S^*} \left[ (W_S^* - w_\mu^*) (W_S^* - w_\mu^*)^\mathbf{T} \right] \right\}.
\end{aligned} \tag{B.15}$$

Denote  $\tilde{\Sigma}_\mu \triangleq \mathbb{E}_{S, W_S^*} \left[ (W_S^* - w_\mu^*) (W_S^* - w_\mu^*)^\mathbf{T} \right] = \mathbb{E}_{W_S^*} [W_S^* W_S^{*\mathbf{T}}] - w_\mu^* w_\mu^{*\mathbf{T}}$ .

Notice that

$$\begin{aligned}
\mathbb{E}_{S, W_S^*} [\Lambda_{W_S^*}] &= \mathbb{E}_{S, W_S^*, W_T} \left[ (W_T - W_S^*) (W_T - W_S^*)^\mathbf{T} \right] \\
&= \mathbb{E}_{W_T} [W_T W_T^\mathbf{T}] - \mathbb{E}_{W_S^*} [W_S^* W_S^{*\mathbf{T}}] \\
&= \mathbb{E}_{W_T} [W_T W_T^\mathbf{T}] - w_\mu^* w_\mu^{*\mathbf{T}} - \left( \mathbb{E}_{W_S^*} [W_S^* W_S^{*\mathbf{T}}] - w_\mu^* w_\mu^{*\mathbf{T}} \right) \\
&= \Lambda_{w_\mu^*} - \tilde{\Sigma}_\mu.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} \mathbb{E}_{S, W_S^*} [\Lambda_{W_S^*}] - \mathbf{I}_d + \Lambda_{w_\mu^*}^{-1} \mathbb{E}_{W_S^*} \left[ (W_S^* - w_\mu^*) (W_S^* - w_\mu^*)^\mathbf{T} \right] \right\} \\
&= \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} \mathbb{E}_{S, W_S^*} [\Lambda_{W_S^*}] - \Lambda_{w_\mu^*}^{-1} \Lambda_{w_\mu^*} + \Lambda_{w_\mu^*}^{-1} \mathbb{E}_{W_S^*} \left[ (W_S^* - w_\mu^*) (W_S^* - w_\mu^*)^\mathbf{T} \right] \right\} \\
&= \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} \left( \mathbb{E}_{S, W_S^*} [\Lambda_{W_S^*}] - \Lambda_{w_\mu^*} + \tilde{\Sigma}_\mu \right) \right\} \\
&= 0.
\end{aligned}$$

Plugging this into Eq. (B.15), we have

$$\begin{aligned}
& \mathbb{E}_{S, W_S^*} \left[ \log \frac{\det(\Lambda_{w_\mu^*})}{\det(\Lambda_{W_S^*})} + \text{tr} \left\{ \Lambda_{w_\mu^*}^{-1} \Lambda_{W_S^*} - \mathbf{I}_d \right\} + (W_S^* - w_\mu^*)^\mathbf{T} \Lambda_{w_\mu^*}^{-1} (W_S^* - w_\mu^*) \right] \\
&= \mathbb{E}_{S, W_S^*} \left[ \log \frac{\det(\Lambda_{w_\mu^*})}{\det(\Lambda_{W_S^*})} \right] = \mathbb{E}_{S, W_S^*} \left[ \text{tr} \left\{ \log \left( \Lambda_{W_S^*}^{-1} \Lambda_{w_\mu^*} \right) \right\} \right].
\end{aligned}$$

Finally, applying Theorem B.3.1 will conclude the proof.  $\square$

### B.3.4 Proof of Corollary 4.4.1

*Proof.* The proof is straightforward by plugging  $\Lambda_{w^*} = \left[ H_{w^*} \left( \frac{2}{\eta} \mathbf{I}_d \right) \right]^{-1} C_T$  in Theorem 4.4.1.  $\square$

### B.3.5 Proof of Corollary 4.4.2

*Proof.* By Lemma B.2.2, it's easy to obtain the following bound according to Theorem 4.4.1.

$$\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2 d}{2n} \log \left( \frac{\mathbb{E} [\mathrm{d}_M^2(W_S^*, w_\mu^*; \mathbb{E}[\Lambda_{W_S^*}])]}{d} + 1 \right)} + \mathbb{E} \left[ \mathrm{tr} \left\{ \log \left( \Lambda_{W_S^*}^{-1} \mathbb{E}[\Lambda_{W_S^*}] \right) \right\} \right].$$

Then, plugging  $\Lambda_{W_S^*} = \frac{\eta}{2b} \mathbf{I}_d$  will conclude the proof.  $\square$

### B.3.6 Corollary B.3.1: Distance to Initialization

**Corollary B.3.1.** Under (i-iii) in Lemma 4.4.1, then  $\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{dR^2}{n} \log \left( \frac{2b}{\eta d} \mathbb{E}[\|W_S^* - W_0\|^2] + 1 \right)}$ .

*Proof.* Notice that  $I(W_T; S) \leq \mathbb{E}_S \mathrm{D}_{\mathrm{KL}}(Q_{W_T|S} \| P_{W_T})$  holds for any  $\sigma > 0$ , then for a given  $\tilde{w}$ , we have

$$\begin{aligned} I(W_T; S) &= \inf_{P_{W_T}} \mathbb{E}_S [\mathrm{D}_{\mathrm{KL}}(Q_{W_T|S} \| P_{W_T})] \\ &\leq \inf_{\sigma} \mathbb{E}_S \left[ \mathrm{D}_{\mathrm{KL}}(P_{W_S^* + \sqrt{\frac{\eta}{2b}} N, W_S^* | S} \| P_{\tilde{w} + \sigma N}) \right] \end{aligned} \quad (\text{B.16})$$

$$\begin{aligned} &= \inf_{\sigma} \mathbb{E}_{S, W_S^*} \left[ \mathrm{D}_{\mathrm{KL}}(P_{W_S^* + \sqrt{\frac{\eta}{2b}} N, | S, W_S^*} \| P_{\tilde{w} + \sigma N}) \right] \\ &= \inf_{\sigma} \frac{1}{2} \mathbb{E}_{S, W_S^*} \left[ \frac{1}{\sigma^2} (W_S^* - \tilde{w})^{\mathrm{T}} (W_S^* - \tilde{w}) + \log \frac{\sigma^{2d}}{(\eta/2b)^d} + \mathrm{tr} \left\{ \frac{\eta}{2b\sigma^2} \mathbf{I}_d \right\} - d \right] \\ &= \frac{1}{2} \inf_{\sigma} \frac{1}{\sigma^2} \mathbb{E}_{S, W_S^*} \left[ \|W_S^* - \tilde{w}\|^2 + \frac{\eta d}{2b} \right] + d \log \sigma^2 + d \log \frac{2b}{\eta} - d \\ &= \frac{1}{2} d \log \left( \frac{2b}{\eta d} \mathbb{E}_{S, W_S^*} [\|W_S^* - \tilde{w}\|^2] + 1 \right), \end{aligned} \quad (\text{B.17})$$



where Eq. (B.16) is by the chain rule of KL divergence, and the optimal

$$\sigma^* = \sqrt{\mathbb{E}_{S, W_S^*} \left[ \|W_S^* - \tilde{w}\|^2/d + \frac{\eta}{2b} \right]}.$$

Let  $\tilde{w} = W_0$  will conclude the proof.  $\square$

Additionally, Corollary B.3.1 can be used to recover a trajectory-based bound.

**Corollary B.3.2.** *Let  $W_T = W_s^*$ ,  $\tilde{w} = 0$  and W.L.O.G, assume  $W_0 = 0$ , then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{dR^2}{n} \log \left( \frac{4bT\eta}{d} \sum_{t=1}^T \mathbb{E} [\|G_t\|^2 + \text{tr}\{C_t\}] + 1 \right)},$$

**Remark B.3.1.** *In Theorem 4.3.1, let  $\tilde{g} = 0$  and by applying Jensen's inequality, we could also let the summation and factor  $T$  move inside the square root. Then the most different part in Corollary B.3.2 is that  $A_2(t)$  is now removed from the bound.*

*Proof.* When  $W_0 = 0$ , we notice that

$$W_T = \sum_{t=1}^T -\eta G_t + \eta N_{C_t},$$

where  $N_{C_t} = C_t^{1/2} N_t$ .

Thus,

$$\|W_T\|^2 = \left\| \sum_{t=1}^T -\eta G_t + \eta N_{C_t} \right\|^2 \leq 2T\eta^2 \sum_{t=1}^T \|G_t\|^2 + \|N_{C_t}\|^2$$

Let  $\tilde{w} = 0$ , recall the bound in Corollary B.3.1 and plugging the inequality above, we have

$$\begin{aligned} \mathcal{E}_\mu(\mathcal{A}) &\leq \sqrt{\frac{R^2}{n} d \log \left( \frac{2b}{\eta d} \mathbb{E}_{S, W_T} [\|W_T - \tilde{w}\|^2] + 1 \right)} \\ &\leq \sqrt{\frac{dR^2}{n} \log \left( 4bT\eta/d \mathbb{E}_{S, W_{0:T-1}, N_{C_{0:t-1}}} \left[ \sum_{t=1}^T \|G_t\|^2 + \|N_{C_t}\|^2 \right] + 1 \right)} \\ &= \sqrt{\frac{dR^2}{n} \log \left( \frac{4bT\eta}{d} \sum_{t=1}^T \mathbb{E}_{S, W_{t-1}} [\|G_t\|^2 + \text{tr}\{C_t\}] + 1 \right)} \end{aligned}$$

This concludes the proof.  $\square$

### B.3.7 Proof of Theorem 4.4.2

*Proof.* Let  $P_{W_T|S_J=s_j} = \mathcal{N}(W_{s_j}^*, \frac{\eta}{2b}\mathbf{I}_d)$ , then

$$\begin{aligned} D_{\text{KL}}(Q_{W_T|S=s} || P_{W_T|S_J=s_j}) &= D_{\text{KL}}(Q_{W_s^* + \sqrt{\frac{\eta}{2b}}N|S=s} || P_{W_{s_j}^* + \sqrt{\frac{\eta}{2b}}N|S_J=s_j}) \\ &\leq D_{\text{KL}}(Q_{W_s^* + \sqrt{\frac{\eta}{2b}}N, W_s^*|S=s} || P_{W_{s_j}^* + \sqrt{\frac{\eta}{2b}}N, W_{s_j}^*|S_J=s_j}) \end{aligned} \quad (\text{B.18})$$

$$\begin{aligned} &= \mathbb{E}_{W_s^*, W_{s_j}^*} \left[ D_{\text{KL}}(Q_{W_s^* + \sqrt{\frac{\eta}{2b}}N|W_s^*, S=s} || P_{W_{s_j}^* + \sqrt{\frac{\eta}{2b}}N|W_{s_j}^*, S_J=s_j}) \right] \\ &= \mathbb{E}_{W_s^*, W_{s_j}^*} \left[ \frac{b}{\eta} ||W_s^* - W_{s_j}^*||^2 \right], \end{aligned} \quad (\text{B.19})$$

where Eq. (B.18) is by the chain rule of KL divergence. Plugging the Eq. (B.19) into Lemma 4.2.1 will obtain the final result.  $\square$

## B.4 Experiment Details and Additional Results

The implementation in this paper is on PyTorch [164], and all the experiments are carried out on NVIDIA Tesla V100 GPUs (32 GB). Most experiment settings follow [223], and the code is also based their implementation, which is available at: <https://github.com/uuujf/MultiNoise>.

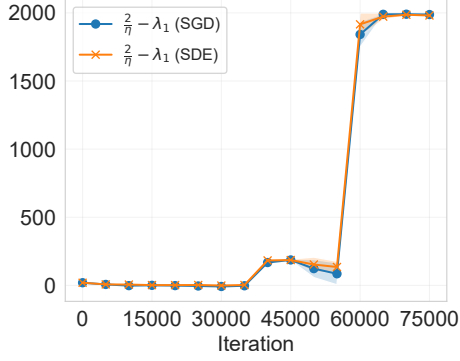
### B.4.1 Hyperparameters

For CIFAR 10, the initial learning rates used for VGG-11 and ResNet-18 are 0.01 and 0.1, respectively. For SVHN, the initial learning rate is 0.05. For CIFAR100, the initial learning rate is 0.1. The learning rate is then decayed by 0.1 at iteration 40,000 and 60,000. If not stated otherwise, the batch size of SGD is 100.

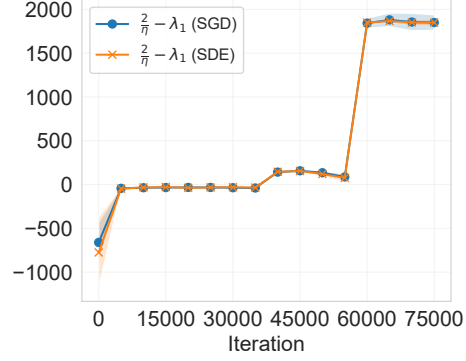
### B.4.2 Additional Empirical Results

## B.5 Additional Result: Inverse Population FIM as both Posterior and Prior Covariance

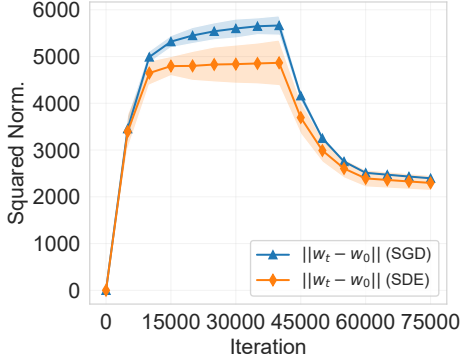
Inspired by some previous works of [1, 81, 212], we can also select the inverse population Fisher information matrix  $F_{w^*}^\mu = \mathbb{E}_Z [\nabla \ell(w^*, Z) \nabla \ell(w^*, Z)^\text{T}]$  as the posterior covariance.



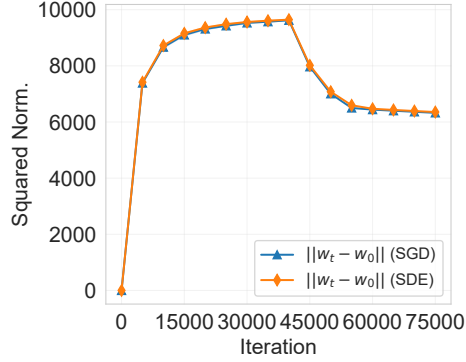
(a) ResNet on CIFAR10



(b) ResNet on CIFAR100



(c) ResNet on CIFAR10



(d) ResNet on CIFAR100

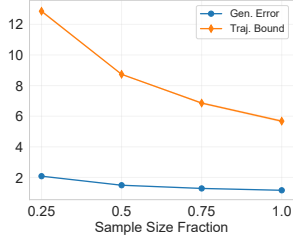
Figure B.1: (a-b) The dynamics of  $\eta/2 - \lambda_1$ . Note that learning rate decays by 0.1 at the 40,000<sup>th</sup> and the 60,000<sup>th</sup> iteration. (c-d) The distance of current model parameters from its initialization.

Then, the following theorem is obtained.

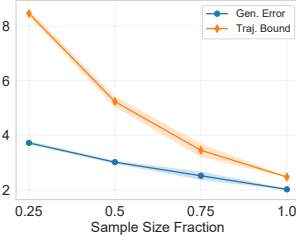
**Theorem B.5.1.** *Under the same conditions in Theorem 4.4.2, and assume the distribution  $P_{W_{S_J}^*|S_J}$  is invariant of  $J$ , then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{M}{2n} \mathbb{E}_S \left[ \sqrt{\mathbb{E}_{W_S^*}^S \left[ \text{tr} \{ H_{W_S^*}^{-1} F_{W_S^*}^\mu \} \right]} \right].$$

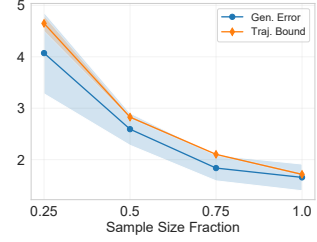
*Proof of Theorem B.5.1.* We now use  $(F_{W_S^*}^\mu)^{-1}$  as both the posterior and prior covariance



(a) VGG on (small) SVHN

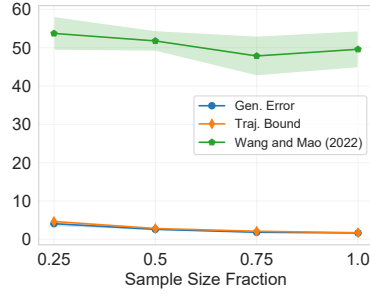


(b) VGG on CIFAR10

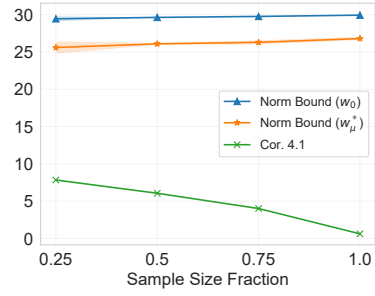


(c) ResNet on CIFAR10

Figure B.2: Zoomed-in of generalization error.



(a) ResNet (Traj. Bound)



(b) ResNet (Term. Bound)

Figure B.3: Estimated trajectories-based bound and terminal-state based bound, with  $R$  excluded. Models trained on CIFAR 10.

(again, we assume  $F_{W_S^*}^\mu \approx F_{W_{S_j}^*}^\mu$  for any  $j$ ), then

$$\begin{aligned}
\mathcal{E}_\mu(\mathcal{A}) &\leq \mathbb{E}_S \left[ \sqrt{\frac{M^2}{4} \mathbb{E}_{J, W_S^*, W_{S_j}^*}^S \left[ (W_S^* - W_{S_j}^*) F_{W_S^*}^\mu (W_S^* - W_{S_j}^*)^\mathbf{T} \right]} \right] \\
&= \frac{M}{2n} \mathbb{E}_S \left[ \sqrt{\mathbb{E}_{W_S^*, W_{S_j}^*}^S \left[ \text{tr} \left\{ F_{W_S^*}^\mu H_{W_S^*}^{-1} H_{W_{S_j}^*}^{-1} \mathbb{E}_J [\nabla \ell(W_S^*, Z_i) \nabla \ell(W_{S_j}^*, Z_i)^\mathbf{T}] \right\} \right]} \right] \\
&= \frac{M}{2n} \mathbb{E}_S \left[ \sqrt{\mathbb{E}_{W_S^*}^S \left[ \text{tr} \left\{ F_{W_S^*}^\mu H_{W_S^*}^{-1} \right\} \right]} \right],
\end{aligned}$$

which completes the proof.  $\square$

**Remark B.5.1.** Notice that  $F_{W_S^*}^\mu \approx H_{W_S^*}^\mu \approx \Sigma^\mu(W_S^*)$  near minima [165, Chapter 8], then  $\text{tr}\{H_{W_S^*}^{-1} \Sigma^\mu(W_S^*)\}$  is very close to the Takeuchi Information Criterion [202]. In addition,

our bound in Theorem [B.5.1](#) is similar to [[196](#), Theorem 3.] with the same convergence rate, although strictly speaking, their result is not a generalization bound. Moreover, as also pointed out in [[196](#)], here  $H_{W_S^*}^{-1}$  is evaluated on the training sample, unlike other works that evaluates the inverse Hessian on the testing sample (e.g., [[203](#)]).

The invariance assumption is also used in [[208](#)]. In practice,  $n$  is usually very large, when  $m = n - 1$ , this assumption indicates that replacing one instance in  $s_j$  will not make  $P_{W_{s_j}^*|s_j}$  be too different.

# Appendix C

## Appendix of Chapter 5

### C.1 Some Additional Definitions on Chaining Methods

Definition C.1.1-C.1.4 are used in the context of chaining methode, e.g., Theorem 5.3.4, Corollary C.2.1 and Theorem C.3.1.

The following is a technique assumption.

**Definition C.1.1** (Separable Process). *The random process  $\{X_t\}_{t \in \mathcal{T}}$  is called separable if there is a countable set  $T_0 \subseteq T$  s.t.  $X_t \in \lim_{s \rightarrow t, s \in T_0} X_s$  for  $\forall t \in T$  a.s., where  $x \in \lim_{s \rightarrow t, s \in T_0} x_s$  means that there is a sequence  $(s_n)$  in  $T_0$  s.t.  $s_n \rightarrow t$  and  $x_{s_n} \rightarrow x$ .*

**Definition C.1.2** (Subgaussian Process). *The random process  $\{X_t\}_{t \in T}$  on the metric space  $(T, d)$  is called subgaussian if  $\mathbb{E}[X_t] = 0$  for all  $t \in T$  and  $\mathbb{E}[e^{\lambda(X_t - X_s)}] \leq e^{\frac{1}{2}\lambda^2 d^2(t, s)}$  for all  $t, s \in T$ ,  $\lambda \geq 0$ .*

**Definition C.1.3** (Stochastic Chain [248]). *Let  $(X_{\mathcal{T}}, T)$  be a random process and random variable pair, where  $T$  is a random variable in the index set  $\mathcal{T}$ . A sequence of random variables  $\{T_k\}_{k=k_0}^{\infty}$  (with each distributed in  $\mathcal{T}$ ) is called a stochastic chain of the pair  $(X_{\mathcal{T}}, T)$ , if 1)  $\lim_{k \rightarrow \infty} \mathbb{E}[X_{T_k}] = \mathbb{E}[X_T]$ , 2)  $\mathbb{E}[X_{T_{k_0}}] = 0$  and 3)  $\{X_t\}_{t \in \mathcal{T}} - T - T_k - T_{k-1}$  is a Markov chain for every  $k > k_0$ .*

**Definition C.1.4** (Increasing Sequence of  $\epsilon$ -Partition). *A partition  $\mathcal{P} = \{A_1, A_2, \dots, A_m\}$  of the set  $T$  is called an  $\epsilon$ -partition of the metric space  $(T, d)$  if for all  $i = 1, 2, \dots, m$ ,  $A_i$  can be contained within a ball of radius  $\epsilon$ . A sequence of partitions  $\{\mathcal{P}_k\}_{k=m}^{\infty}$  of a set  $T$  is called an increasing sequence if for any  $k \geq m$  and each  $A \in \mathcal{P}_{k+1}$ , there exists  $B \in \mathcal{P}_k$  s.t.  $A \subseteq B$ .*

## C.2 Omitted Proofs and Additional Results in Section 5.3

### C.2.1 Proof of Theorem 5.3.1

The following proof shows that the proof of e-CMI bound in [92] can be adapted to the loss-difference MI bound, where we just replace the distribution  $P_{L_i^+, L_i^-}$  by  $P_{\Delta L_i}$ .

*Proof.* Let  $U'_i$  be the independent copy of  $U_i$  s.t.  $U'_i \sim \text{Bern}(1/2)$  and  $U' \perp\!\!\!\perp \Delta L_i | \tilde{Z}$ . Recall Lemma 2.4.1,

$$\begin{aligned} I(\Delta L_i; U_i | \tilde{Z} = \tilde{z}) &= \text{D}_{\text{KL}} \left( P_{\Delta L_i, U_i | \tilde{Z} = \tilde{z}} \| P_{\Delta L_i | \tilde{Z} = \tilde{z}} P_{U'_i} \right) \\ &\geq \sup_{t \in \mathbb{R}} \mathbb{E}_{\Delta L_i, U_i | \tilde{z}} [tg(\Delta L_i, U_i, \tilde{z})] - \ln \mathbb{E}_{\Delta L_i, U'_i | \tilde{z}} \left[ e^{tg(\Delta L_i, U'_i, \tilde{z})} \right]. \end{aligned}$$

Recall that  $w = \mathcal{A}(\tilde{z}_u)$ , we now let  $g(\Delta l_i, u_i, \tilde{z}) = (-1)^{u_i} \Delta l_i = (-1)^{u_i} (\ell(\mathcal{A}(\tilde{z}_u), \tilde{z}_i^-) - \ell(\mathcal{A}(\tilde{z}_u), \tilde{z}_i^+))$ . Then,

$$\begin{aligned} I(\Delta L_i; U_i | \tilde{Z} = \tilde{z}) &\geq \sup_t \mathbb{E}_{\Delta L_i, U_i | \tilde{z}} [t(-1)^{U_i} \Delta L_i] - \ln \mathbb{E}_{\Delta L_i, U'_i | \tilde{z}} \left[ e^{t(-1)^{U'_i} \Delta L_i} \right] \\ &= \sup_t \mathbb{E}_{\Delta L_i, U_i | \tilde{z}} [t(-1)^{U_i} \Delta L_i] - \ln \mathbb{E}_{\Delta L_i | \tilde{z}} \left[ \mathbb{E}_{U'} \left[ e^{t(-1)^{U'_i} \Delta L_i} | \Delta L_i = \Delta l_i \right] \right], \end{aligned} \tag{C.1}$$

where the last equality is by the conditional independence.

Since  $\mathbb{E}_{U'} [t(-1)^{U'_i} \Delta l_i] = 0$  and  $(-1)^{U'_i}$  is bounded between  $[-1, 1]$ , by Lemma 2.4.2, we have

$$\mathbb{E}_{U'_i} \left[ e^{t(-1)^{U'_i} \Delta l_i} \right] \leq e^{\frac{t^2 \Delta l_i^2}{2}} \leq e^{\frac{t^2}{2}},$$

where the second inequality is by the boundedness condition of the loss function, i.e.  $\Delta l_i \in [-1, 1]$ .

Plugging the inequality above into Eq. (C.1), we have

$$I(\Delta L_i; U_i | \tilde{Z} = \tilde{z}) \geq \sup_t \mathbb{E}_{\Delta L_i, U_i | \tilde{z}} [t(-1)^{U_i} \Delta L_i] - \frac{t^2}{2}.$$

Then consider the case of  $t > 0$  and  $t < 0$  ( $t = 0$  is trivial), by AM–GM inequality (i.e. the arithmetic mean is greater than or equal to the geometric mean), the following is straightforward,

$$|\mathbb{E}_{\Delta L_i, U_i | \tilde{Z}} [(-1)^{U_i} \Delta L_i]| \leq \sqrt{2I(\Delta L_i; U_i | \tilde{Z} = \tilde{z})}.$$

Notice that

$$|\mathcal{E}_\mu(\mathcal{A})| = |\mathbb{E}_{S,W} [L_\mu(W) - L_S(W)]| = |\mathbb{E}_{\tilde{Z}, U, W} [L_{\tilde{Z} \setminus \tilde{Z}_U}(W) - L_{\tilde{Z}_U}(W)]| \quad (\text{C.2})$$

$$\leq \mathbb{E}_{\tilde{Z}} |\mathbb{E}_{U, W | \tilde{Z}} [L_{\tilde{Z}_U}(W) - L_{\tilde{Z}_U}(W)]| \quad (\text{C.3})$$

$$\begin{aligned} &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}} |\mathbb{E}_{U_i, W | \tilde{Z}} [(-1)^{U_i} (\ell(W, \tilde{Z}_i^-) - \ell(W, \tilde{Z}_i^+))]| \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}} |\mathbb{E}_{U_i, \Delta L_i | \tilde{Z}} [(-1)^{U_i} \Delta L_i]|, \end{aligned}$$

wherein the two inequalities are by applying the Jensen's inequality to the absolute function.

Hence, putting everything together we have

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}} \sqrt{2I^{\tilde{Z}}(\Delta L_i; U_i)} \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2I(\Delta L_i; U_i | \tilde{Z})},$$

where the second inequality is by applying the Jensen's inequality to the square root function.

This completes the proof.  $\square$

### C.2.2 Proof of Theorem 5.3.2

By revisiting the proof of Theorem 5.3.1, particularly Eq. (C.2-C.3), we notice that if we do not move the expectation over  $\tilde{Z}$  outside of the absolute function, we will have a chance to get ride of the expectation over  $\tilde{Z}$  if we take the expectation over  $\Delta L_i$ .



*Proof.* By the definition of the expected generalization error, we have

$$\begin{aligned} |\mathcal{E}_\mu(\mathcal{A})| &= |\mathbb{E}_{S,W} [L_\mu(W) - L_S(W)]| = \left| \mathbb{E}_{\tilde{Z},U,W} \left[ L_{\tilde{Z} \setminus \tilde{Z}_U}(W) - L_{\tilde{Z}_U}(W) \right] \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z},U_i,W} \left[ (-1)^{U_i} \left( \ell(W, \tilde{Z}_i^-) - \ell(W, \tilde{Z}_i^+) \right) \right] \right| \end{aligned} \quad (\text{C.4})$$

$$\leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{\Delta L_i, U_i} [(-1)^{U_i} \Delta L_i] \right|. \quad (\text{C.5})$$

We know that  $(-1)^{U_i} \Delta L_i$  is bounded between  $[-1, 1]$ , so it is a 1-subgaussian random variable. Then, recall Lemma 2.4.5 and let  $g(X, Y) = (-1)^{U_i} \Delta L_i$ , we have

$$\left| \mathbb{E}_{\Delta L_i, U_i} [(-1)^{U_i} \Delta L_i] - \mathbb{E}_{\Delta L_i, U'_i} [(-1)^{U'_i} \Delta L_i] \right| \leq \sqrt{2I(\Delta L_i; U_i)}.$$

Since  $\mathbb{E}_{\Delta L_i, U'_i} [(-1)^{U'_i} \Delta L_i] = 0$ , plugging the inequality above into Eq. (C.5), we have

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{\Delta L_i, U_i} [(-1)^{U_i} \Delta L_i] \right| \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2I(\Delta L_i; U_i)}.$$

This concludes the proof.  $\square$

### C.2.3 Proof of Theorem 5.3.5

*Proof.* Recall Eq. (C.5), we could also obtain

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{\Delta L_i, U_i} [(-1)^{U_i} \Delta L_i] \right| = \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{\Delta L_i, U_i} [(-1)^{U_i} \Delta L_i] - \mathbb{E}_{\Delta L'_i, U_i} [(-1)^{U_i} \Delta L'_i] \right|,$$

where  $\Delta L'_i$  is an independent copy of  $\Delta L_i$  (i.e.  $\Delta L'_i \sim P_{\Delta L_i}$  and  $\Delta L'_i \perp\!\!\!\perp U_i$ ) and the second equality holds since  $\mathbb{E}_{\Delta L'_i, U_i} [(-1)^{U_i} \Delta L'_i] = 0$ .

Then, by Jensen's inequality, we move the expectation over  $U_i$  and the average outside the absolute function,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{U_i} \left[ \left| \mathbb{E}_{\Delta L_i | U_i} [(-1)^{U_i} \Delta L_i] - \mathbb{E}_{\Delta L'_i} [(-1)^{U_i} \Delta L'_i] \right| \middle| U_i = u_i \right].$$

Notice that for any fixed  $u_i$ , we have

$$\begin{aligned}\mathbb{E}_{\Delta L_i|U_i=u_i} [(-1)^{u_i} \Delta L_i] &= \int_{-1}^1 (-1)^{u_i} \Delta \ell_i dP_{\Delta L_i|U_i=u_i}(\Delta \ell_i), \\ \mathbb{E}_{\Delta L'_i} [(-1)^{u_i} \Delta L'_i] &= \int_{-1}^1 (-1)^{u_i} \Delta \ell'_i dP_{\Delta L_i}(\Delta \ell'_i).\end{aligned}$$

Also, noting that  $f(x) = x$  is a 1-Lipschitz function, i.e.  $|(-1)^{u_i} \Delta L_i - (-1)^{u_i} \Delta L'_i| \leq |\Delta L_i - \Delta L'_i|$  holds trivially.

Recall the KR duality of Wasserstein distance (i.e. Lemma 2.4.7), we have

$$\begin{aligned}|\mathcal{E}_\mu(\mathcal{A})| &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{U_i} \left[ \left| \mathbb{E}_{\Delta L_i|U_i} [(-1)^{U_i} \Delta L_i] - \mathbb{E}_{\Delta L'_i} [(-1)^{U_i} \Delta L'_i] \right| \middle| U_i = u_i \right] \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{U_i} [\mathbb{W}(P_{\Delta L_i|U_i}, P_{\Delta L_i})].\end{aligned}$$

This concludes the proof.  $\square$

### C.2.4 Proof of Theorem 5.3.3

*Proof.* The channel capacity can be obtained from Lemma C.4.1 by letting  $\epsilon_i = 0$  and changing the unit of bit to the unit of nat (i.e. replacing logarithm base of 2 to the base of  $e$  by  $\ln x = \ln 2 \log_2 x$ ).

Furthermore, the value of  $1 - \alpha_i$  reflects the chance that the interpolating learning algorithm  $\mathcal{A}$  will make an error (i.e.  $\ell(W, Z'_i) = 1$ ) for a testing instance, or equivalently,

$$\begin{aligned}\mathbb{E}_{W, Z'_i} [\mathbb{1}_{f_W(X'_i) \neq Y'_i}] &= \mathbb{E}_{U_i, L_i} [L_{i, \bar{U}_i}] = \frac{\mathbb{E}_{L_i^-|U_i=0} [L_i^-] + \mathbb{E}_{L_i^+|U_i=1} [L_i^+]}{2} \\ &= \frac{P(\Delta L_i = 1|U_i = 0) + P(\Delta L_i = -1|U_i = 1)}{2} = 1 - \alpha_i.\end{aligned}$$

Thus, combining the equality above with  $C = I(U_i; \Delta L_i) = (1 - \alpha_i) \cdot \ln 2$ , we have

$$|\mathcal{E}_\mu(\mathcal{A})| = \mathbb{E}_W [L_\mu(W)] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W, Z'_i} [\mathbb{1}_{f_W(X'_i) \neq Y'_i}] = \frac{1}{n} \sum_{i=1}^n (1 - \alpha_i) = \frac{1}{n \ln 2} \sum_{i=1}^n I(U_i; \Delta L_i).$$

This completes the proof.  $\square$

### C.2.5 Proof of Theorem 5.3.4

*Proof.* Let  $E_{\Delta L_i}^i = \mathcal{E}_\mu(\mathcal{A})^i(\Delta L_i) = (-1)^{U_i} \Delta L_i$ , then for any integers  $k_1$  and  $k_0$  such that  $k_1 > k_0$ , we have

$$E_{\Delta L_i}^i = E_{\Delta L_{i,k_0}}^i + \sum_{k=k_0+1}^{k_1} (E_{\Delta L_{i,k}}^i - E_{\Delta L_{i,k-1}}^i) + E_{\Delta L_i}^i - E_{\Delta L_{i,k_1}}^i.$$

By the definition of the stochastic chain (i.e. Definition C.1.3), we know that  $\mathbb{E}_{E_{\Delta L_{i,k_0}}^i} [E_{\Delta L_{i,k_0}}^i] = 0$  and  $\lim_{k_1 \rightarrow \infty} E_{\Delta L_{i,k_1}}^i = E_{\Delta L_i}^i$ . Therefore, let  $k_1 \rightarrow \infty$  and taking expectation over  $(U_i, \Delta L_i) \sim P_{U_i, \Delta L_i}$  for both sides of the equation above, we have

$$\mathbb{E}_{U_i, \Delta L_i} [E_{\Delta L_i}^i] = \sum_{k=k_0+1}^{\infty} \mathbb{E}_{U_i, \Delta L_{i,k}, \Delta L_{i,k-1}} [E_{\Delta L_{i,k}}^i - E_{\Delta L_{i,k-1}}^i]. \quad (\text{C.6})$$

Let  $U'_i$  be an independent copy of  $U_i$  and recall Lemma 2.4.1, we have

$$\begin{aligned} & \mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} [\text{D}_{\text{KL}}(P_{U_i | \Delta L_{i,k}, \Delta L_{i,k-1}} || P_{U'_i})] \\ & \geq \mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} \left[ \sup_{t>0} t \mathbb{E}_{U_i | \Delta L_{i,k}, \Delta L_{i,k-1}} [E_{\Delta L_{i,k}}^i - E_{\Delta L_{i,k-1}}^i] - \ln \mathbb{E}_{U'_i} \left[ e^{t(E_{\Delta L_{i,k}}^i - E_{\Delta L_{i,k-1}}^i)} \right] \right] \\ & \geq \sup_{t>0} t \mathbb{E}_{U_i, \Delta L_{i,k}, \Delta L_{i,k-1}} [E_{\Delta L_{i,k}}^i - E_{\Delta L_{i,k-1}}^i] - \mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} \ln \mathbb{E}_{U'_i} \left[ e^{t(-1)^{U'_i}(\Delta L_{i,k} - \Delta L_{i,k-1})} \right], \end{aligned}$$

where the second inequality is by applying Jensen's inequality to the supremum.

Notice that the LHS above is equivalent to  $I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i)$ . Since  $(-1)^{U'_i}$  is bounded between  $[-1, 1]$ , by Lemma 2.4.2, we have

$$I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i) \geq \sup_{t>0} t \mathbb{E}_{U_i, \Delta L_{i,k}, \Delta L_{i,k-1}} [E_{\Delta L_{i,k}}^i - E_{\Delta L_{i,k-1}}^i] - \mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} \ln e^{\frac{t^2(\Delta L_{i,k} - \Delta L_{i,k-1})^2}{2}}.$$

Thus, let  $d(\Delta L_{i,k}, \Delta L_{i,k-1}) = |\Delta L_{i,k} - \Delta L_{i,k-1}|$ , we have

$$\mathbb{E}_{U_i, \Delta L_{i,k}, \Delta L_{i,k-1}} [E_{\Delta L_{i,k}}^i - E_{\Delta L_{i,k-1}}^i] \leq \sqrt{2 \mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} [d^2(\Delta L_{i,k}, \Delta L_{i,k-1})] I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i)}.$$

Plugging the inequality above into Eq. (C.6) and taking average over  $i$ , we have,

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{U_i, \Delta L_i} [E_{\Delta L_i}^i] \\ &\leq \frac{1}{n} \sum_{i=1}^n \sum_{k=k_0+1}^{\infty} \sqrt{2\mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} [d^2(\Delta L_{i,k}, \Delta L_{i,k-1})] I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i)}.\end{aligned}$$

From the third point of Definition C.1.3, we know that  $U_i - \Delta L_i - \Delta L_{i,k} - \Delta L_{i,k-1}$  is a Markov chain, so  $I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i) = I(\Delta L_{i,k}; U_i) + I(\Delta L_{i,k-1}; U_i | \Delta L_{i,k}) = I(\Delta L_{i,k}; U_i)$ . This gives us the final form of the bound,

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{1}{n} \sum_{i=1}^n \sum_{k=k_0+1}^{\infty} \sqrt{2\mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} [d^2(\Delta L_{i,k}, \Delta L_{i,k-1})] I(\Delta L_{i,k}; U_i)}.$$

This concludes the proof.  $\square$

## C.2.6 Additional Result: Chained MI Bound for Bounded Loss

**Corollary C.2.1.** *Let  $2^{-k_0} \geq \text{diam}(\Gamma)$  and let  $\{\mathcal{P}_k\}_{k=k_0}^{\infty}$  be an increasing sequence of partitions of  $\Gamma$ , where for each  $k \geq k_0$ ,  $\mathcal{P}_k$  is a  $2^{-k}$ -partition of  $(\Gamma, d)$ . Let  $\Delta L_{i,k}$  be the center of the covering ball of the partition cell that  $\Delta L_i$  belongs to the partition  $\mathcal{P}_k$ , then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{3}{n} \sum_{i=1}^n \sum_{k=k_0}^{\infty} 2^{-k} \sqrt{2I(\Delta L_{i,k}; U_i)}.$$

*Proof.* By the triangle inequality, we have  $d(\Delta L_{i,k}, \Delta L_{i,k-1}) \leq d(\Delta L_{i,k}, \Delta L_i) + d(\Delta L_i, \Delta L_{i,k-1})$ .

Since  $\mathcal{P}_k$  is a  $2^{-k}$  partition,  $d(\Delta L_{i,k}, \Delta L_i) \leq 2^{-k}$ , then  $d(\Delta L_{i,k}, \Delta L_i) + d(\Delta L_i, \Delta L_{i,k-1}) \leq 2^{-k} + 2^{-(k-1)} = 3 \times 2^{-k}$ . Plugging this into Theorem 5.3.4, we have

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) &\leq \frac{1}{n} \sum_{i=1}^n \sum_{k=k_0+1}^{\infty} \sqrt{2\mathbb{E}_{\Delta L_{i,k}, \Delta L_{i,k-1}} [d^2(\Delta L_{i,k}, \Delta L_{i,k-1})] I(\Delta L_{i,k}; U_i)} \\ &\leq \frac{1}{n} \sum_{i=1}^n \sum_{k=k_0+1}^{\infty} 3 \times 2^{-k} \sqrt{2I(\Delta L_{i,k}; U_i)}.\end{aligned}$$

This completes the proof.  $\square$

## C.3 Omitted Proofs and Additional Results in Section 5.4

### C.3.1 Proof of Lemma 5.4.1

*Proof.* By the definition of  $\mathcal{E}_\mu(\mathcal{A})$ , we can decompose it into two terms,

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, L_i^-, U_i} [L_{i, \bar{U}_i} - L_{i, U_i}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, L_i^-, U_i} [(-1)^{U_i} (L_i^- - L_i^+)] \\ &= \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{L_i^-, U_i} [(-1)^{U_i} L_i^-] + \mathbb{E}_{L_i^+, U_i} [-(-1)^{U_i} L_i^+] \right] \\ &= \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{L_i^-, U_i} [(-1)^{U_i} L_i^-] + \mathbb{E}_{L_i^+, \bar{U}_i} [(-1)^{\bar{U}_i} L_i^+] \right],\end{aligned}$$

where the last equality is by  $-(-1)^{U_i} L_i^+ = (-1)^{\bar{U}_i} L_i^+$ . We now show that the following holds

$$\mathbb{E}_{L_i^-, U_i} [(-1)^{U_i} L_i^-] = \mathbb{E}_{L_i^+, \bar{U}_i} [(-1)^{\bar{U}_i} L_i^+].$$

Recall that  $\tilde{Z}$  and  $U$  are i.i.d drawn from  $\mu^{2n}$  and the Bernoulli distribution, respectively, and  $\tilde{Z} \perp\!\!\!\perp U$ . Usually, a learning algorithm may depend on the order of training instances (i.e. for  $i \neq j$ , even if two training instances satisfy  $z_i = z_j$ ,  $P_{W|z_i}$  may not be the same with  $P_{W|z_j}$ ), but it should be invariant to the *row index* of the supersample  $\tilde{Z}$ , then the distribution  $P_{L_i^-, U_i}$  and the distribution  $P_{L_i^+, U_i}$  have some symmetric property, namely,  $P_{L_i^-|U_i=1} = P_{L_i^+|U_i=0}$  and  $P_{L_i^-|U_i=0} = P_{L_i^+|U_i=1}$ . Or equivalently, the distribution of the training loss (or testing loss) of the  $i$ th training instance (or testing instance) is invariant to  $U_i$ . Hence, we have  $P_{L_i^-} = P_{L_i^+}$ , we say  $L_i^-$  and  $L_i^+$  are identically but not independently distributed. Then,

$$\begin{aligned}\mathbb{E}_{L_i^-, U_i=0} [L_i^-] &= \int_0^1 \ell_i^- dP_{L_i^-|U_i=0}(\ell_i^-) = \int_0^1 \ell_i^+ dP_{L_i^+|U_i=1}(\ell_i^+) = \mathbb{E}_{L_i^+|U_i=1} [L_i^+], \\ \mathbb{E}_{L_i^-, U_i=1} [-L_i^-] &= \int_0^1 -\ell_i^- dP_{L_i^-|U_i=1}(\ell_i^-) = \int_0^1 -\ell_i^+ dP_{L_i^+|U_i=0}(\ell_i^+) = \mathbb{E}_{L_i^+|U_i=0} [-L_i^+].\end{aligned}$$

These give us

$$\begin{aligned}\mathbb{E}_{L_i^-, U_i} [(-1)^{U_i} L_i^-] &= \frac{\mathbb{E}_{L_i^-|U_i=0} [L_i^-] + \mathbb{E}_{L_i^-|U_i=1} [-L_i^-]}{2} = \frac{\mathbb{E}_{L_i^+|U_i=1} [L_i^+] + \mathbb{E}_{L_i^+|U_i=0} [-L_i^+]}{2} \\ &= \mathbb{E}_{L_i^+, \bar{U}_i} [(-1)^{\bar{U}_i} L_i^+].\end{aligned}$$

Therefore,

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) &= \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{L_i^-, U_i} \left[ (-1)^{U_i} L_i^- \right] + \mathbb{E}_{L_i^+, \bar{U}_i} \left[ (-1)^{\bar{U}_i} L_i^+ \right] \right] = \frac{2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^-, U_i} \left[ (-1)^{U_i} L_i^- \right] \\ &= \frac{2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \bar{U}_i} \left[ (-1)^{\bar{U}_i} L_i^+ \right].\end{aligned}$$

Notice that both  $(-1)^{U_i}$  and  $(-1)^{\bar{U}_i}$  are Rademacher variables. This conclude the proof.  $\square$

### C.3.2 Proof of Theorem 5.4.1

*Proof.* Notice that  $2(-1)^{U'_i} L_i^+$  is bounded between  $[-2, 2]$ , it is a subgaussian random variable with the variance proxy  $\sigma = 2$ . Let the measurable function  $g(L_i^+, U_i)$  in Lemma 2.4.5 be  $2(-1)^{U_i} L_i^+$ , then  $g(L_i^+, U_i)$  is 2-subgaussian under  $P_{U_i} P_{L_i^+}$ , we have

$$\left| 2\mathbb{E}_{L_i^+, U_i} \left[ (-1)^{U_i} L_i^+ \right] - 2\mathbb{E}_{L_i^+, U'_i} \left[ (-1)^{U'_i} L_i^+ \right] \right| \leq 2\sqrt{2I(L_i^+; U_i)}.$$

Since  $\mathbb{E}_{L_i^+, U'_i} \left[ (-1)^{U'_i} L_i^+ \right] = \frac{\mathbb{E}_{L_i^+} [L_i^+] - \mathbb{E}_{L_i^+} [L_i^+]}{2} = 0$ , then

$$\left| 2\mathbb{E}_{L_i^+, U_i} \left[ (-1)^{U_i} L_i^+ \right] \right| \leq 2\sqrt{2I(L_i^+; U_i)}.$$

Recall Lemma 5.4.1,

$$\mathcal{E}_\mu(\mathcal{A}) = \frac{2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \bar{U}_i} \left[ (-1)^{\bar{U}_i} L_i^+ \right].$$

Notice that  $\bar{U}_i$  and  $U_i$  are one-to-one mapping, so using any of them will give the same mutual information. Thus, by applying Jensen's inequality to the absolute function, we have

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \left| 2\mathbb{E}_{L_i^+, U_i} \left[ (-1)^{U_i} L_i^+ \right] \right| \leq \frac{2}{n} \sum_{i=1}^n \sqrt{2I(L_i^+; U_i)}.$$

In addition, to obtain the second inequality in the theorem, we first invoke the independence between  $\tilde{Z}$  and  $U_i$ ,  $I(L_i^+; U_i) + I(\tilde{Z}; U_i | L_i^+) = I(L_i^+, \tilde{Z}; U_i) = I(L_i^+; U_i | \tilde{Z})$  and

then use the DPI,  $I(L_i^+; U_i | \tilde{Z}) \leq I(f_W(X); U_i | \tilde{Z})$ , we have

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{2}{n} \sum_{i=1}^n \sqrt{2I(L_i^+; U_i)} \leq \frac{2}{n} \sum_{i=1}^n \sqrt{2I(f_W(X_i^+); U_i | \tilde{Z})}.$$

This completes the proof.  $\square$

### C.3.3 Proof of Theorem 5.4.2

*Proof.* Notice that

$$\begin{aligned} \mathbb{E}_{W, Z'_i} [\mathbb{1}_{f_W(X'_i) \neq Y'}] &= \mathbb{E}_{U_i, L_i} [L_i, \bar{U}_i] \\ &= \frac{\mathbb{E}_{L_i^- | U_i=0} [L_i^-] + \mathbb{E}_{L_i^+ | U_i=1} [L_i^+]}{2} \\ &= \mathbb{E}_{L_i^+ | U_i=1} [L_i^+] \\ &= P(L_i^+ = 1 | U_i = 1) \\ &= 1 - q_i. \end{aligned}$$

Hence,  $L_\mu = \sum_{i=1}^n \frac{1-q_i}{n}$ .

For each  $i$ , we have  $I(L_i^+; U_i) = H(L_i^+) - H(L_i^+ | U_i) = H(\frac{1-q_i}{2}) - \frac{1}{2}H(1 - q_i) \leq H(\frac{1-q_i}{2})$ . Since the entropy function  $H(\cdot)$  is a concave function, we have  $\frac{1}{n} \sum_{i=1}^n I(L_i^+; U_i) \leq \frac{1}{n} \sum_{i=1}^n H(\frac{1-q_i}{2}) \leq H(\frac{L_\mu}{2})$ .  $\square$

### C.3.4 Additional Result: Chained MI Bound Based on Single-Loss

When the loss is not discrete or even not bounded, let  $\xi^i(\ell_i^+) \triangleq \varepsilon_i \ell_i^+$  be a random process and let  $\mathcal{L}$  be the domain of  $L_i^+$ . Similar to Theorem 5.3.4, we can also have the corresponding chained bound of Theorem 5.4.1.

**Theorem C.3.1.** *For each  $i \in [n]$ , we assume  $\{L_{i,k}^+\}_{k=k_0}^\infty$  is a stochastic chain of  $(\xi^i(\ell_i^+))_{\ell_i^+ \in \mathcal{L}, L_i^+}$ , then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{2}{n} \sum_{i=1}^n \sum_{k=k_0}^\infty \sqrt{2\mathbb{E} [d^2(L_{i,k}^+, L_{i,k-1}^+)] I(L_{i,k}^+; U_i)},$$

where the RHS expectation is taken over  $(L_{i,k}^+, L_{i,k-1}^+)$ .

This theorem can be obtained by following the same development with the proof in Section C.2.5.

### C.3.5 Proof of Lemma 5.4.2

*Proof.* A key step is the second equality of the following

$$\begin{aligned}
\mathcal{E}_\mu^{C_1}(\mathcal{A}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{L_i^-, L_i^+, U_i} [L_{i, \bar{U}_i} - (1 + C_1)L_{i, U_i}] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{L_i^-, L_i^+, U_i} \left[ \left(1 + \frac{C_1}{2}\right) (L_{i, \bar{U}_i} - L_{i, U_i}) - \frac{C_1}{2} L_{i, \bar{U}_i} - \frac{C_1}{2} L_{i, U_i} \right] \\
&= \frac{2 + C_1}{2n} \sum_{i=1}^n \left[ \mathbb{E}_{L_i^-, U_i} \left[ (-1)^{U_i} L_i^- - \frac{C_1}{C_1 + 2} L_i^- \right] + \mathbb{E}_{L_i^+, U_i} \left[ -(-1)^{U_i} L_i^+ - \frac{C_1}{C_1 + 2} L_i^+ \right] \right].
\end{aligned}$$

Recall that  $P_{L_i^-} = P_{L_i^+}$ , we have  $\mathbb{E}_{L_i^-} \left[ \frac{C_1}{C_1 + 2} L_i^- \right] = \mathbb{E}_{L_i^+} \left[ \frac{C_1}{C_1 + 2} L_i^+ \right]$ . Also, noting that  $\mathbb{E}_{L_i^- | U_i=0} [L_i^-] = \mathbb{E}_{L_i^+ | U_i=1} [L_i^+]$  and  $\mathbb{E}_{L_i^- | U_i=1} [-L_i^-] = \mathbb{E}_{L_i^+ | U_i=0} [-L_i^+]$ , we have

$$\mathcal{E}_\mu^{C_1}(\mathcal{A}) = \frac{2 + C_1}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \bar{U}_i} \left[ (-1)^{\bar{U}_i} L_i^+ - \frac{C_1}{C_1 + 2} L_i^+ \right] = \frac{2 + C_1}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i} [\tilde{\varepsilon}_i L_i^+], \quad (\text{C.7})$$

where  $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1}{C_1 + 2}$  and  $\varepsilon_i \sim \text{Unif}\{-1, +1\}$  is the Rademacher variable. In this case,  $\tilde{\varepsilon}_i$  is called the *shifted* Rademacher variable and its mean is  $-\frac{C_1}{C_1 + 2}$ .  $\square$

### C.3.6 Proof of Theorem 5.4.3

*Proof.* Since  $\tilde{\varepsilon}_i$  is obtained by a bijection function of  $U_i$ , they two can be replaced by each other in the mutual information. Recall Lemma 2.4.1 and let the measurable function  $g$  be  $t(C_1 + 2)\tilde{\varepsilon}_i L_i^+$ , we have

$$\begin{aligned}
I(L_i^+; U_i) &= I(L_i^+; \tilde{\varepsilon}_i) = \text{D}_{\text{KL}} \left( P_{L_i^+, \tilde{\varepsilon}_i} || P_{L_i^+} P_{\tilde{\varepsilon}_i'} \right) \\
&\geq \sup_{t > 0} \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i} \left[ t(C_1 + 2)\tilde{\varepsilon}_i L_i^+ \right] - \ln \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i'} \left[ e^{t(C_1 + 2)\tilde{\varepsilon}_i' L_i^+} \right]. \quad (\text{C.8})
\end{aligned}$$

We hope to have

$$\mathbb{E}_{L_i^+, \tilde{\varepsilon}_i'} \left[ e^{t(C_1 + 2)\tilde{\varepsilon}_i' L_i^+} \right] \leq 1. \quad (\text{C.9})$$



Since  $\tilde{\varepsilon}'_i$  is independent of  $L_i^+$ , and  $P(\tilde{\varepsilon}_i = \frac{2}{C_1+2}) = P(\tilde{\varepsilon}_i = \frac{-2(C_1+1)}{C_1+2}) = \frac{1}{2}$ , then

$$\mathbb{E}_{L_i^+, \tilde{\varepsilon}'_i} \left[ e^{t(C_1+2)\tilde{\varepsilon}'_i L_i^+} \right] = \frac{\mathbb{E}_{L_i^+} \left[ e^{-2t(C_1+1)L_i^+} + e^{2tL_i^+} \right]}{2}.$$

Notice that  $e^{-2t(C_1+1)L_i^+} + e^{2tL_i^+}$  is the summation of two convex function, which is still a convex function, so the maximum value of this function is achieved at the endpoints of the bounded domain. Recall that  $L_i^+ \in [0, 1]$ , we now consider two cases, i) when  $L_i^+ = 0$ , we have  $e^{-2t(C_1+1)L_i^+} + e^{2tL_i^+} = 2$ ; ii) when  $L_i^+ = 1$ , we need to require  $e^{-2t(C_1+1)} + e^{2t} \leq 2$  s.t. Eq (C.9) can hold. Note that this inequality implies that  $t \leq \frac{\ln 2}{2}$ .

Replacing  $t$  by  $C_2$ , let the values of  $C_1, C_2$  be taken from the domain of  $\{C_1, C_2 | C_1, C_2 > 0, e^{-2C_2(C_1+1)} + e^{2C_2} \leq 2\}$ , so Eq. (C.9) will hold. Under this condition, by re-arranging the terms in Eq. (C.8), we have

$$(C_1 + 2)\mathbb{E}_{L_i^+, \tilde{\varepsilon}_i} [\tilde{\varepsilon}_i L_i^+] \leq \frac{I(L_i^+; U_i)}{C_2}.$$

Plugging the inequality above into Eq. (C.7), we have

$$\mathcal{E}_\mu^{C_1}(\mathcal{A}) = L_\mu - (1 + C_1)L_n = \frac{2 + C_1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i} [\tilde{\varepsilon}_i L_i^+] \right] \leq \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n}.$$

Thus, the following inequality can be obtained,

$$L_\mu \leq \min_{C_1, C_2 > 0, e^{2C_2} + e^{-2C_2(C_1+1)} \leq 2} (1 + C_1)L_n + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n}. \quad (\text{C.10})$$

We can also optimize the parameters  $C_1, C_2$  by relaxing the condition of  $e^{-2C_2(C_1+1)} + e^{2C_2} \leq 2$ . By invoking  $e^x \geq x + 1$  and  $e^{-x} \leq \frac{1}{1+x}$  for  $x > -1$ , and  $e^x \leq \frac{1}{1-x}$  for  $x < 1$ , it's sufficient to have

$$\frac{1}{1 + 2C_2(C_1 + 1)} + \frac{1}{1 - 2C_2} \leq 2, \quad \text{and } 0 < C_2 < \frac{1}{2}. \quad (\text{C.11})$$

Solving Eq (C.11) gives us  $0 < C_2 \leq \frac{C_1}{4(C_1+1)}$ . Since  $\frac{C_1}{4(C_1+1)} \leq \frac{1}{4} < \frac{1}{2}$ , then Eq (C.11) holds when  $C_2 \in (0, \frac{C_1}{4(C_1+1)}]$ . Notice that  $\frac{C_1}{4(C_1+1)}$  is also smaller than  $\frac{\ln 2}{2}$ .

Therefore, we obtain

$$\begin{aligned}
L_\mu &\leq \min_{C_1, C_2 > 0, e^{2C_2} + e^{-2C_2(C_1+1)} \leq 2} (1 + C_1)L_n + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n} \\
&\leq \min_{C_1 > 0, 0 < C_2 \leq \frac{C_1}{4(C_1+1)}} (1 + C_1)L_n + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{C_2 n} \\
&= L_n + \sum_{i=1}^n \frac{4I(L_i^+; U_i)}{n} + 4 \sqrt{\sum_{i=1}^n \frac{L_n I(L_i^+; U_i)}{n}},
\end{aligned}$$

where the last equality is achieved when  $C_1 = 2\sqrt{\sum_{i=1}^n \frac{I(L_i^+; U_i)}{nL_n}}$  and  $C_2 = \frac{C_1}{4(C_1+1)}$ .

For the second part of the theorem, if  $\mathcal{A}$  is an interpolating algorithm, then  $L_n = 0$ , in which case we can let  $C_1$  be arbitrarily large.

Recall that we hope

$$e^{-2C_2(C_1+1)} + e^{2C_2} \leq 2.$$

This can be satisfied by letting  $C_2 = \frac{\ln 2}{2}$  and  $C_1 \rightarrow \infty$ .

Thus, the interpolating single-loss MI bound is

$$L_\mu \leq \sum_{i=1}^n \frac{2I(L_i^+; U_i)}{n \ln 2}.$$

This completes the proof. □

### C.3.7 Proof of Lemma 5.4.3

*Proof.* By the definition of  $\gamma$ -variance, and notice that  $L_S(W) = \frac{1}{n} \sum_{i=1}^n \ell(W, Z_i)$ , we have

$$\begin{aligned}
V(\gamma) &= \mathbb{E}_{W,S} \left[ \frac{1}{n} \sum_{i=1}^n (\ell(W, Z_i) - (1 + \gamma)L_S(W))^2 \right] \\
&= \mathbb{E}_{W,S} \left[ \frac{1}{n} \sum_{i=1}^n (\ell^2(W, Z_i) - 2(1 + \gamma)\ell(W, Z_i)L_S(W) + (1 + \gamma)^2 L_S^2(W)) \right] \\
&= \mathbb{E}_{W,S} \left[ \frac{1}{n} \sum_{i=1}^n \ell^2(W, Z_i) \right] - \mathbb{E}_{W,S} [(1 - \gamma^2)L_S^2(W)] \\
&= L_n - (1 - \gamma^2)\mathbb{E}_{W,S} [L_S^2(W)],
\end{aligned}$$

where the last equality is due to the fact that the loss is the zero-one loss (i.e.  $\ell^2(\cdot, \cdot) = \ell(\cdot, \cdot)$ ).  $\square$

### C.3.8 Proof of Lemma 5.4.4

*Proof.* By Lemma 5.4.3 and  $\gamma \in (0, 1)$ , we notice that,

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) - C_1 V(\gamma) &= L_\mu - L_n - C_1 L_n + C_1(1 - \gamma^2) \mathbb{E}_{W,S} [L_S^2(W)] \\ &\leq L_\mu - (1 + C_1) L_n + C_1(1 - \gamma^2) \mathbb{E}_{W,S} [L_S(W)]\end{aligned}\tag{C.12}$$

$$= L_\mu - (1 + C_1 \gamma^2) L_n,\tag{C.13}$$

where the inequality is because that  $L_S(W) \in [0, 1]$  (i.e.  $L_S^2(W) \leq L_S(W)$ ).

Noting that in Eq (C.13),  $\mathcal{E}_\mu(\mathcal{A}) - C_1 V(\gamma)$  is upper bounded by a weighted generalization error. Thus, we can then directly apply Lemma 5.4.2 by choosing  $C_1 \gamma^2$  as the trade-off coefficient (i.e. replacing  $C_1$  in Lemma 5.4.2 by  $C_1 \gamma^2$ ), which gives us

$$\mathcal{E}_\mu(\mathcal{A}) - C_1 V(\gamma) \leq \frac{2 + C_1 \gamma^2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i} [\tilde{\varepsilon}_i L_i^+],$$

where  $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1 \gamma^2}{C_1 \gamma^2 + 2}$ .  $\square$

### C.3.9 Proof of Theorem 5.4.4

*Proof.* The RHS of Eq. (C.13) in the proof of Lemma 5.4.4 has already been bounded in Theorem 5.4.3 by regarding  $C_1 \gamma^2$  as the weighted parameter  $C_1$  in Theorem 5.4.3. Then, there exist  $C_1, C_2 > 0$  s.t.

$$\mathcal{E}_\mu(\mathcal{A}) - C_1 V(\gamma) \leq L_\mu - (1 + C_1 \gamma^2) L_n \leq \sum_{i=1}^n \frac{I(L_i^+; U_i)}{n C_2}.$$

Furthermore, from the proof of Theorem 5.4.3, we note that the following is valid

$$\mathcal{E}_\mu(\mathcal{A}) \leq \min_{C_1, C_2 > 0, e^{2C_2} + e^{-2C_2(C_1 \gamma^2 + 1)} \leq 2} C_1 V(\gamma) + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{n C_2}.$$

Notice that the original optimization space of the variance based bound should be larger than  $\{C_1, C_2 | C_1, C_2 > 0, e^{2C_2} + e^{-2C_2(C_1 \gamma^2 + 1)} \leq 2\}$  because in Eq. (C.12), we upper bound the most interested quantity  $\mathcal{E}_\mu(\mathcal{A}) - C_1 V(\gamma)$  by  $L_\mu - (1 + C_1 \gamma^2) L_n$ , which restricts the original optimization space. This completes the proof.  $\square$

### C.3.10 Proof of Lemma 5.4.5

*Proof.* By the definition of  $\lambda$ -sharpness, we have

$$\begin{aligned} F_i(\lambda) &= \mathbb{E}_{Z_i} [\mathbb{E}_{W|Z_i} [\ell(W, Z_i)^2] - 2(1 + \lambda)\mathbb{E}_{W|Z_i} \ell(W, Z_i) + (1 + \lambda)^2 \mathbb{E}_{W|Z_i}^2 \ell(W, Z_i)] \\ &= \mathbb{E}_{W, Z_i} [\ell(W, Z_i)] - (1 - \lambda^2) \mathbb{E}_{Z_i} [\mathbb{E}_{W|Z_i}^2 \ell(W, Z_i)], \end{aligned}$$

where the last equality is by the zero-one loss. This completes the proof.  $\square$

### C.3.11 Proof of Lemma 5.4.6

*Proof.* By Lemma 5.4.5, we have

$$\begin{aligned} \mathcal{E}_\mu(\mathcal{A}) - \frac{C_1}{n} \sum_{i=1}^n F_i(\lambda) &= L_\mu - (1 + C_1)L_n + \frac{(1 - \lambda^2)C_1}{n} \sum_{i=1}^n \mathbb{E}_{Z_i} [\mathbb{E}_{W|Z_i}^2 \ell(W, Z_i)] \\ &= \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{U_i, L_i} [L_{i, \bar{U}_i} - (1 + C_1)L_{i, U_i}] + (1 - \lambda^2)C_1 \mathbb{E}_{\tilde{Z}_i, U_i} \left[ \mathbb{E}_{L_i, U_i | \tilde{Z}_i, U_i}^2 L_{i, U_i} \right] \right]. \end{aligned}$$

Let  $\Lambda(\tilde{Z}_i, U_i) = \mathbb{E}_{L_i, U_i | \tilde{Z}_i, U_i}^2 L_{i, U_i}$ ,  $\Lambda(\tilde{Z}_i, \bar{U}_i) = \mathbb{E}_{L_i, \bar{U}_i | \tilde{Z}_i, \bar{U}_i}^2 L_{i, \bar{U}_i}$ ,  $\Lambda(\tilde{Z}_i^+) = \mathbb{E}_{L_i^+ | \tilde{Z}_i^+, U_i}^2 L_i^+$  and  $\Lambda(\tilde{Z}_i^-) = \mathbb{E}_{L_i^- | \tilde{Z}_i^-, U_i}^2 L_i^-$ . A key observation is the following:

$$\begin{aligned} & \mathbb{E}_{U_i, L_i} [L_{i, \bar{U}_i} - (1 + C_1)L_{i, U_i}] + (1 - \lambda^2)C_1 \mathbb{E}_{\tilde{Z}_i, U_i} [\Lambda(\tilde{Z}_i, U_i)] \\ &= (1 + \frac{C_1}{2}) \mathbb{E}_{U_i, L_i} [L_{i, \bar{U}_i} - L_{i, U_i}] - \frac{C_1}{2} (1 - \lambda^2) \mathbb{E}_{\tilde{Z}_i, U_i} [\Lambda(\tilde{Z}_i, \bar{U}_i) - \Lambda(\tilde{Z}_i, U_i)] \\ & \quad - \frac{C_1}{2} \left[ \mathbb{E}_{U_i, L_i} [L_{i, \bar{U}_i}] - (1 - \lambda^2) \mathbb{E}_{\tilde{Z}_i, U_i} [\Lambda(\tilde{Z}_i, \bar{U}_i)] \right] - \frac{C_1}{2} \left[ \mathbb{E}_{U_i, L_i} [L_{i, U_i}] - (1 - \lambda^2) \mathbb{E}_{\tilde{Z}_i, U_i} [\Lambda(\tilde{Z}_i, U_i)] \right] \\ &= \mathbb{E}_{L_i^-, U_i} \left[ (-1)^{U_i} \frac{C_1 + 2}{2} L_i^- - (-1)^{U_i} \frac{C_1(1 - \lambda^2)}{2} \mathbb{E}_{\tilde{Z}_i^- | U_i} [\Lambda(\tilde{Z}_i^-)] \right. \\ & \quad \left. - \frac{C_1}{2} L_i^- + \frac{C_1(1 - \lambda^2)}{2} \mathbb{E}_{\tilde{Z}_i^- | U_i} [\Lambda(\tilde{Z}_i^-)] \right] \\ & \quad + \mathbb{E}_{L_i^+, U_i} \left[ -(-1)^{U_i} \frac{C_1 + 2}{2} L_i^+ + (-1)^{U_i} \frac{C_1(1 - \lambda^2)}{2} \mathbb{E}_{\tilde{Z}_i^+ | U_i} [\Lambda(\tilde{Z}_i^+)] \right. \\ & \quad \left. - \frac{C_1}{2} L_i^+ + \frac{C_1(1 - \lambda^2)}{2} \mathbb{E}_{\tilde{Z}_i^+ | U_i} [\Lambda(\tilde{Z}_i^+)] \right] \\ &= (C_1 + 2) \mathbb{E}_{L_i^+, U_i} \left[ (\varepsilon_i - \frac{C_1}{C_1 + 2}) L_i^+ - \frac{C_1(1 - \lambda^2)}{C_1 + 2} (\varepsilon_i - 1) \mathbb{E}_{\tilde{Z}_i^+ | U_i} [\Lambda(\tilde{Z}_i^+)] \right], \end{aligned}$$

where  $\varepsilon_i$  is the Rademacher variable.

Thus,

$$\begin{aligned} & \mathcal{E}_\mu(\mathcal{A}) - \frac{C_1}{n} \sum_{i=1}^n F_i(\lambda) \\ &= \frac{C_1 + 2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, U_i} \left[ \left( \varepsilon_i - \frac{C_1}{C_1 + 2} \right) L_{i,1} - \frac{C_1(1 - \lambda^2)}{C_1 + 2} (\varepsilon_i - 1) \mathbb{E}_{\tilde{Z}_i^+ | U_i} \left[ \mathbb{E}_{L_i^+ | \tilde{Z}_i^+, U_i}^2 L_i^+ \right] \right]. \end{aligned} \quad (\text{C.14})$$

This completes the proof.  $\square$

### C.3.12 Proof of Theorem 5.4.5

*Proof.* Notice that we cannot directly apply Lemma 2.4.1 starting from here since there is a quadratic term, namely,  $\mathbb{E}_{\tilde{Z}_i^+ | U_i} \left[ \mathbb{E}_{L_i^+ | \tilde{Z}_i^+, U_i}^2 L_i^+ \right]$  in the RHS.

Inspired by [233], we now assume that there exists a random variable  $R_i$  s.t.

$$\begin{aligned} & (C_1 + 2) \mathbb{E}_{L_i^+, U_i} \left[ \left( \varepsilon_i - \frac{C_1}{C_1 + 2} \right) L_i^+ - \frac{C_1(1 - \lambda^2)}{C_1 + 2} (\varepsilon_i - 1) \mathbb{E}_{\tilde{Z}_i^+ | U_i} \left[ \mathbb{E}_{L_i^+ | \tilde{Z}_i^+, U_i}^2 L_i^+ \right] \right] \\ & \leq (C_1 + 2) \mathbb{E}_{L_i^+, U_i} \left[ \left( \varepsilon_i - \frac{C_1}{C_1 + 2} \right) L_i^+ - \frac{C_1(1 - \lambda^2)}{C_1 + 2} (\varepsilon_i - 1) \mathbb{E}_{\tilde{Z}_i^+ | U_i} \left[ R_i \mathbb{E}_{L_i^+ | \tilde{Z}_i^+, U_i} L_i^+ \right] \right] \\ & = (C_1 + 2) \mathbb{E}_{L_i^+, U_i} \left[ \left( \varepsilon_i - \frac{C_1}{C_1 + 2} \right) L_i^+ - \frac{C_1(1 - \lambda^2)}{C_1 + 2} (\varepsilon_i - 1) \mathbb{E}_{L_i^+ | U_i} \left[ R_i L_i^+ \right] \right] \\ & = (C_1 + 2) \mathbb{E}_{L_i^+, U_i} \left[ \left( \left( \varepsilon_i - \frac{C_1}{C_1 + 2} \right) - \frac{C_1(1 - \lambda^2)}{C_1 + 2} (\varepsilon_i - 1) R_i \right) L_i^+ \right]. \end{aligned} \quad (\text{C.15})$$

Such  $R_i$  could satisfy  $R_i \geq \sup_{\tilde{Z}_i^+} \mathbb{E}_{L_i^+ | \tilde{Z}_i^+ = \tilde{Z}_i^+, U_i = u_i} L_i^+$ , for any fixed  $u_i$ , and the randomness of  $R_i$  is controlled by  $U_i$ , i.e.  $R_i$  is a function of  $U_i$ . A simple choice is to let  $R_i = 1$  (so  $R_i$  always exists), and another choice could be letting  $R_i = \mathbb{E}_{L_i^{+'} | U_i \sim Q_i} [L_i^{+'}]$  that satisfies the condition, where  $Q_i$  is some distribution of  $L_i^+$ , and  $L_i^{+'}$  is independent of  $L_i^+$  and  $\tilde{Z}_i^+$  given  $U_i$ .

Recall that the shifted Rademacher variable  $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1}{C_1 + 2}$ , and let another shifted Rademacher variable  $\hat{\varepsilon}_i = \varepsilon_i - 1$ . Then we are ready to invoke Lemma 2.4.1,

$$I(L_i^+; U_i) \geq \sup_{t > 0} t \mathbb{E}_{L_i^+, U_i} \left[ \left( (C_1 + 2) \tilde{\varepsilon}_i - C_1(1 - \lambda^2) \hat{\varepsilon}_i R_i \right) L_i^+ \right] - \ln \mathbb{E}_{L_i^+, U_i} \left[ e^{t \left( (C_1 + 2) \tilde{\varepsilon}_i' - C_1(1 - \lambda^2) \hat{\varepsilon}_i' R_i' \right) L_i^+} \right]. \quad (\text{C.16})$$

Similar to the proof of Theorem 5.4.3, we hope the following hold

$$\mathbb{E}_{L_i^+} \mathbb{E}_{U_i'} \left[ e^{t((C_1+2)\tilde{\varepsilon}'_i - C_1(1-\lambda^2)\tilde{\varepsilon}'_i R'_i) L_i^+} \right] \leq 1. \quad (\text{C.17})$$

By the independence, we have

$$\mathbb{E}_{L_i^+} \mathbb{E}_{U_i'} \left[ e^{t((C_1+2)\tilde{\varepsilon}'_i - C_1(1-\lambda^2)\tilde{\varepsilon}'_i R'_i) L_i^+} \right] = \mathbb{E}_{L_i^+} \left[ \frac{e^{2t(C_1(1-\lambda^2)\tilde{r}'_i - C_1 - 1)L_i^+} + e^{2tL_i^+}}{2} \right],$$

where  $\tilde{r}'_i \in [0, 1]$  is the value (or the realization) of  $R'_i$  when  $U'_i = 0$  (or  $\varepsilon'_i = -(-1)^{U'_i} = -1$ ), e.g.,  $\tilde{r}'_i = \mathbb{E}_{L_i^{+'} | U'_i=0} [L_i^{+'}]$ .

Then, since  $L_i^+$  could be either 0 or 1. We now consider the two cases.

(i) When  $L_i^+ = 0$ , then  $\frac{e^{2t(C_1(1-\lambda^2)\tilde{r}'_i - C_1 - 1)L_i^+} + e^{2tL_i^+}}{2} = 1$ . Therefore, when  $L_i^+ = 0$ , the value of  $R'_i$  has no effect on the moment generating function.

(ii) When  $L_i^+ = 1$ , we have the formula  $\frac{e^{2t(C_1(1-\lambda^2)\tilde{r}'_i - C_1 - 1)L_i^+} + e^{2tL_i^+}}{2}$ . Notably, only when  $\varepsilon'_i = -1$  (or  $U'_i = 0$ ) and  $L_i^+ = 1$ , the value of  $R'_i$ , viz,  $\tilde{r}'_i$ , has some impact on the moment generating function. Since  $\tilde{r}'_i \in [0, 1]$ , it's sufficient to upper bound  $R'_i$  by the random variable  $\frac{1-\varepsilon'_i}{2} = \frac{-\tilde{\varepsilon}'_i}{2}$ . Thus,

$$\mathbb{E}_{L_i^+} \mathbb{E}_{U_i'} \left[ e^{t((C_1+2)\tilde{\varepsilon}'_i - C_1(1-\lambda^2)\tilde{\varepsilon}'_i R'_i) L_i^+} \right] \leq \mathbb{E}_{L_i^+} \mathbb{E}_{U_i'} \left[ e^{t((C_1+2)\tilde{\varepsilon}'_i + \frac{C_1}{2}(1-\lambda^2)\tilde{\varepsilon}'_i{}^2) L_i^+} \right]. \quad (\text{C.18})$$

By the moment generating function of the Bernoulli random variable  $L_i^+$ , we have

$$\begin{aligned} \mathbb{E}_{U_i'} \mathbb{E}_{L_i^+} \left[ e^{t((C_1+2)\tilde{\varepsilon}'_i + \frac{C_1}{2}(1-\lambda^2)\tilde{\varepsilon}'_i{}^2) L_i^+} \right] &= \mathbb{E}_{U_i} \left[ 1 - \mathbb{E}_{L_i^+} [L_i^+] + \mathbb{E}_{L_i^+} [L_i^+] e^{t((C_1+2)\tilde{\varepsilon}'_i + \frac{C_1}{2}(1-\lambda^2)\tilde{\varepsilon}'_i{}^2)} \right] \\ &= 1 - \mathbb{E}_{L_i^+} [L_i^+] + \mathbb{E}_{L_i^+} [L_i^+] \frac{e^{-2t(C_1\lambda^2+1)} + e^{2t}}{2}. \end{aligned}$$

Since  $0 \leq \mathbb{E}_{L_i^+} [L_i^+] \leq 1$ , we only need to require that

$$\frac{e^{-2t(C_1\lambda^2+1)} + e^{2t}}{2} \leq 1.$$

Replacing  $t$  by  $C_2$  and putting everything together (Eq. (C.14-C.17)), we have

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{C_1}{n} \sum_{i=1}^n F_i(\lambda) + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{nC_2}.$$

This completes the proof.  $\square$

### C.3.13 Proof of Corollary 5.4.1

*Proof.* According to the proofs in Section C.3.9 and Section C.3.12, we know that the sufficient conditions to let Eq. (5.5) and Eq. (5.6) hold are  $e^{2C_2} + e^{-2C_2(C_1\gamma^2+1)} \leq 2$  and  $e^{2C_2} + e^{-2C_2(C_1\lambda^2+1)} \leq 2$ , respectively. Given that both  $\gamma, \lambda \in (0, 1)$ , there must exist  $C_1, C_2$  to let both Eq. (5.5) and Eq. (5.6) hold. Then taking minimum of these two inequalities will give us the desired result.  $\square$

Additionally, in any of the following case: (i)  $L_n \rightarrow 0$ ; (ii)  $V(\gamma) \rightarrow 0$  for some  $\gamma \in (0, 1)$ ; (iii)  $F(\lambda) \rightarrow 0$  for some  $\lambda \in (0, 1)$ , we can let  $C_1 \rightarrow \infty$  and let  $C_2 = \frac{\ln 2}{2}$ , then  $\mathcal{E}_\mu(\mathcal{A}) \leq \sum_{i=1}^n \frac{2I(L_i^+; U_i)}{n \ln 2}$ . This justifies the remark after Corollary 5.4.1.

## C.4 Some Background on Channel Capacity of Binary Channel

In this section, we follow the custom of the notations in [41], where the logarithm usually has a base of 2 (i.e.  $\log_2$ ). In addition, for a binary random variable, the entropy function  $H(\cdot)$  can be a binary entropy function, for example, the random variable  $X$  has the value 0 and 1, and  $P(X = 1) = p$ ,  $P(X = 0) = 1 - p$ , then  $H(X) = H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p)$ . The channel capacity of a channel between input  $X$  and output  $Y$  is defined as  $C \triangleq \max_{P_X} I(X; Y)$ .

### C.4.1 Channel Capacity of Binary Symmetric Channel (BSC)

In a general case, the channel capacity of Figure 5.1(left) can be computed as in the following lemma.

**Lemma C.4.1.** *When  $X \sim P_{U_i}$ , the channel capacity of the channel in Figure 5.1(left) is achieved and  $C = (1 - \alpha) \left(1 - H\left(\frac{1-\alpha-\epsilon}{1-\alpha}, \frac{\epsilon}{1-\alpha}\right)\right)$ .*

We note that this lemma is an exercise problem in [41, Problem 7.13].

*Proof.* Let  $P(X = 0) = \pi$  and  $P(X = 1) = 1 - \pi$ , then  $I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(1 - \epsilon - \alpha, \alpha, \epsilon)$ . It's easy to see that  $H(Y) = H(\pi(1 - \alpha - \epsilon) + \epsilon(1 - \pi), \alpha, \pi\epsilon +$

$(1 - \pi)(1 - \epsilon - \alpha)$ ). We let  $A = \pi(1 - \alpha - \epsilon) + \epsilon(1 - \pi)$  and  $B = \pi\epsilon + (1 - \pi)(1 - \epsilon - \alpha)$ . Notice that  $A + B = 1 - \alpha$ , then

$$\begin{aligned}
H(Y) &= H(\pi(1 - \alpha - \epsilon) + \epsilon(1 - \pi), \alpha, \pi\epsilon + (1 - \pi)(1 - \epsilon - \alpha)) \\
&= - \left[ (A + B) \log_2(1 - \alpha) + \alpha \log_2 \alpha + A \log_2 \frac{A}{1 - \alpha} + B \log_2 \frac{B}{1 - \alpha} \right] \\
&= H(\alpha) - (1 - \alpha) \left[ \frac{A}{1 - \alpha} \log_2 \frac{A}{1 - \alpha} + \frac{B}{1 - \alpha} \log_2 \frac{B}{1 - \alpha} \right] \\
&= H(\alpha) + (1 - \alpha) H\left(\frac{A}{1 - \alpha}, \frac{B}{1 - \alpha}\right) \leq H(\alpha) + 1 - \alpha.
\end{aligned}$$

To achieve the channel capacity (or to let the equality above hold), we need to let  $A = B$ , which indicates that  $\pi = \frac{1}{2}$ .

Thus,

$$\begin{aligned}
C &= H(\alpha) + 1 - \alpha - H(1 - \epsilon - \alpha, \alpha, \epsilon) \\
&= H(\alpha) + 1 - \alpha - \left( H(\alpha) + (1 - \alpha) H\left(\frac{1 - \epsilon - \alpha}{1 - \alpha}, \frac{\epsilon}{1 - \alpha}\right) \right) \\
&= (1 - \alpha) \left( 1 - H\left(\frac{1 - \alpha - \epsilon}{1 - \alpha}, \frac{\epsilon}{1 - \alpha}\right) \right),
\end{aligned}$$

which completes the proof.  $\square$

## C.5 Experimental Details and Additional Results

### C.5.1 Channel Capacity of Binary Asymmetric Channel (BAC)

The channel capacity of the BAC channel in Figure 5.1(right) is given below.

**Lemma C.5.1.** *The channel capacity of the BAC in Figure 5.1(right) is  $C = \log_2(1 + \beta) - \frac{1-q}{1-p-q}H(p) + \frac{p}{1-p-q}H(q)$ , where  $\beta = 2^{\frac{H(p)-H(q)}{1-p-q}}$  and the capacity is achieved when  $P(U_i = 1) = \frac{1-q(1+\beta)}{(1-p-q)(1+\beta)}$ . Further, if  $p = 0$  (i.e. Z-channel),  $C = \log_2\left(1 + 2^{\frac{-H(q)}{1-q}}\right)$ , and for small  $q$ , the capacity can be approximated by  $C \approx 1 - \frac{1}{2}H(q)$ .*



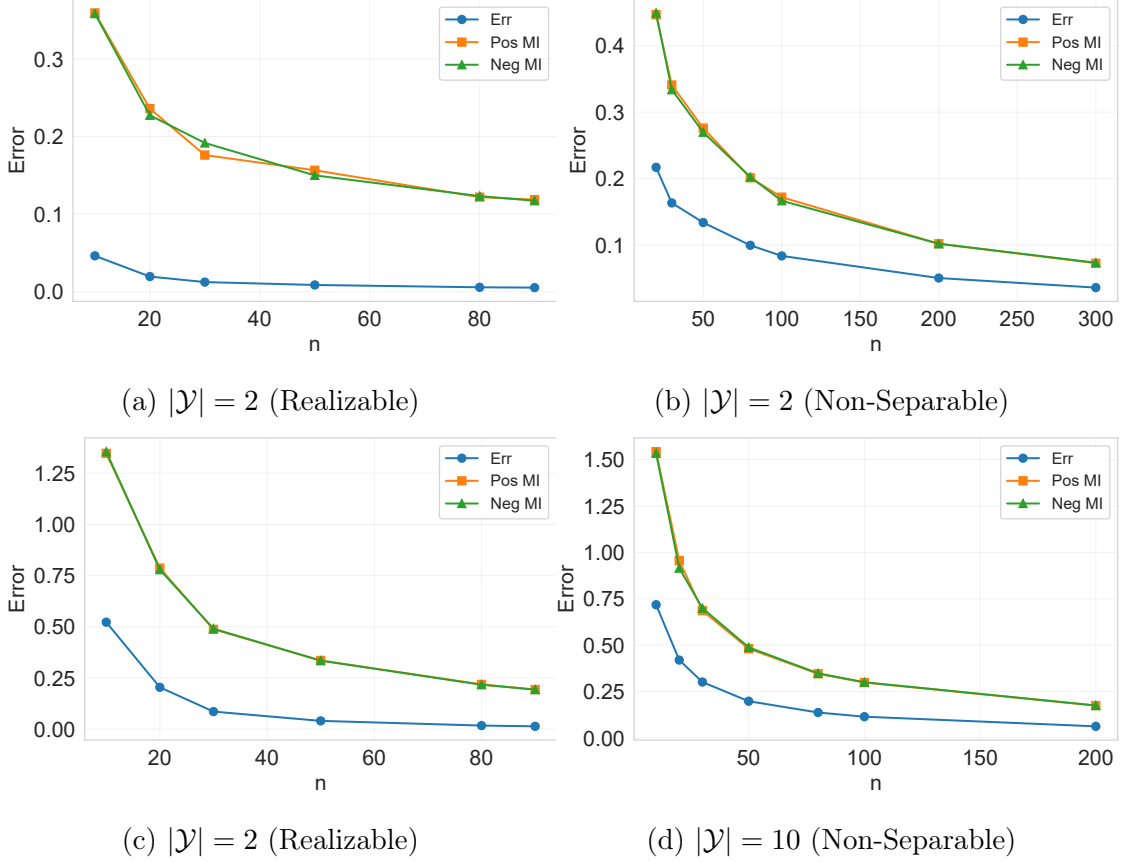


Figure C.1: Comparison of  $I(L_i^+; U_i)$  and  $I(L_i^-; U_i)$ .

**Remark C.5.1.** Notice that in this case, let  $X \sim \text{Bern}(1/2)$  (i.e.  $X = U_i$ ) will not achieve the channel capacity. Thus, in the interpolating setting, except for Theorem 5.4.2, we have another upper bound for  $I(L_i^+; U_i)$ , namely  $\frac{1}{n} \sum_{i=1}^n I(L_i^+; U_i) \leq \frac{1}{n} \sum_{i=1}^n \ln \left( 1 + 2^{\frac{-H(1-q_i)}{1-q_i}} \right)$ . If we further let  $q_i$  be the same for each  $i$  (which indeed should be true), then  $I(L_i^+; U_i) \leq \ln \left( 1 + 2^{\frac{-H(L_\mu)}{L_\mu}} \right)$ .

*Proof.* Let  $P(X = 1) = \pi$ , then  $I(X; Y) = H(\pi(1-p-q) + q) - \pi(H(p) - H(q)) - H(q)$ . Let  $\frac{dI(X; Y)}{d\pi} = (1-p-q) \log_2 \left( \frac{1}{\pi(1-p-q)+q} - 1 \right) - H(p) + H(q) = 0$ , we have the optimal  $\pi^* = \frac{1-q(1+\beta)}{(1-p-q)(1+\beta)}$  where  $\beta = 2^{\frac{H(p)-H(q)}{1-p-q}}$ . Plugging  $\pi = \pi^*$  into the formula of  $I(X; Y)$ , we

have  $I(X; Y) = \log_2(1 + \beta) - \frac{1-q}{1-p-q}H(p) + \frac{p}{1-p-q}H(q)$ . □

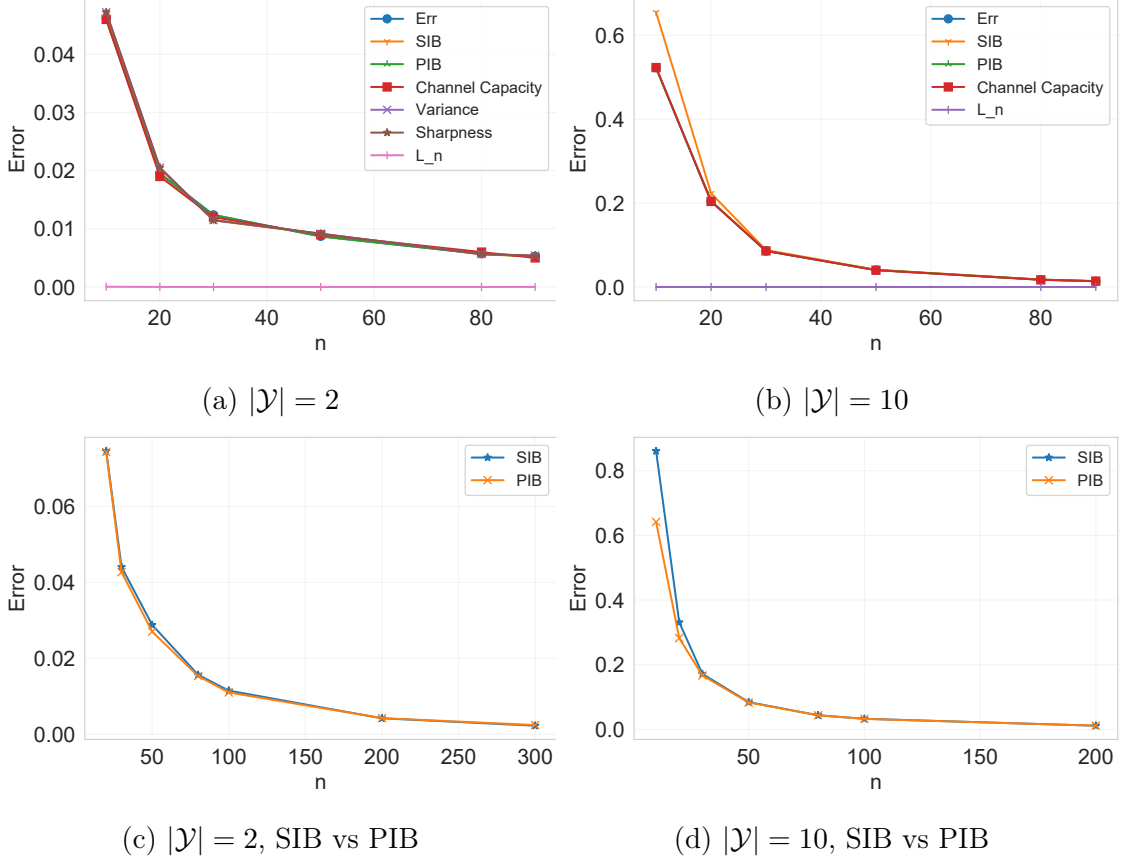


Figure C.2: (a)(b) Comparison of the interpolating bounds on the synthetic dataset. Here *Channel Capacity*, *SIB* and *PIB* refer to the channel capacity result in Theorem 5.3.3, the single-loss based interpolating bound of Eq. (5.3) in Theorem 5.4.3 and a improved loss pair based interpolating bound of [92, Eq. (12)] (replacing  $I(L_i; U_i | \tilde{Z})$  by  $I(L_i; U_i)$ ), respectively. (c)(d) Further compare the value of  $2I(L_i^+; U_i)$  and  $I(L_i; U_i)$ , where training loss is non-zero.

## C.5.2 Experiment Setup

In our linear classifier experiment, we generate synthetic Gaussian data using the widely-used Python package *scikit-learn* [166]. We draw each dimension (or feature) of  $X$  indepen-

dently from some Gaussian distribution, and let all the features be informative to its class labels  $Y$ . Specifically, we choose the dimension of data  $X$  to be 5 and we create different class of points normally distributed (with the standard deviation being 1) about vertices of an 5-dimensional hypercube, where its sides of length can be manually controlled. In addition, we utilize full-batch gradient descent with a fixed learning rate of 0.01 to train the linear classifier. We perform training for a total of 500 epochs, and we employ early stopping when the training error reaches a sufficiently low threshold (e.g.,  $< 0.5\%$ ). To ensure robustness and statistical significance, we draw 50 different supersamples for each experiment. Within each supersample, we further generate 100 different mask random variables, resulting in a total of 5,000 runs for each experimental setting. This comprehensive setup enables us to compare both the unconditional MI bounds and the disintegrated MI bounds. Additionally, if the unconditional MI bound is the sole evaluated objective, one has the option to completely restart the training process 5,000 times.

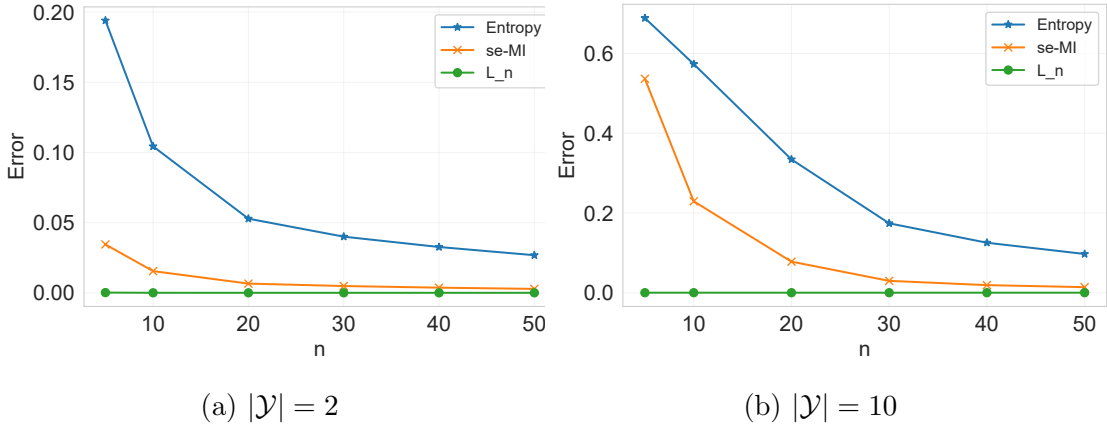


Figure C.3: Verification of Theorem 5.4.2.

In the neural networks experiments, we follow the same setup with [82, 92]. Specifically, we draw  $k_1$  samples of  $\tilde{Z}$  and  $k_2$  samples of  $U$  for each given  $\tilde{z}$ . For the CNN on the binary MNIST dataset, we set  $k_1 = 5$  and  $k_2 = 30$ . The 4-layer CNN model is trained using the Adam optimizer with a learning rate of 0.001 and a momentum coefficient of  $\beta_1 = 0.9$ . The training process spans 200 epochs, with a batch size of 128. For ResNet-50 on CIFAR10, we set  $k_1 = 2$  and  $k_2 = 40$ . The ResNet model is trained using stochastic gradient descent (SGD) with a learning rate of 0.01 and a momentum coefficient of 0.9 for a total of 40 epochs. The batch size for this experiment is set to 64. In the SGLD experiment, we once again train a 4-layer CNN on the binary MNIST dataset. The batch size is set to 100, and the training lasts for 40 epochs. The initial learning rate is 0.01 and decays by a factor

of 0.9 after every 100 iterations. Let  $t$  be the iteration index, the inverse temperature of SGLD is given by  $\min\{4000, \max\{100, 10e^{t/100}\}\}$ . We set the training sample size to  $n = 4000$ , and  $k_1 = 5$  and  $k_2 = 30$ . We save checkpoints every 4 epochs. All these experiments are conducted using NVIDIA Tesla V100 GPUs with 32 GB of memory. For more comprehensive details, including model architectures, we recommend referring to [82, 92].

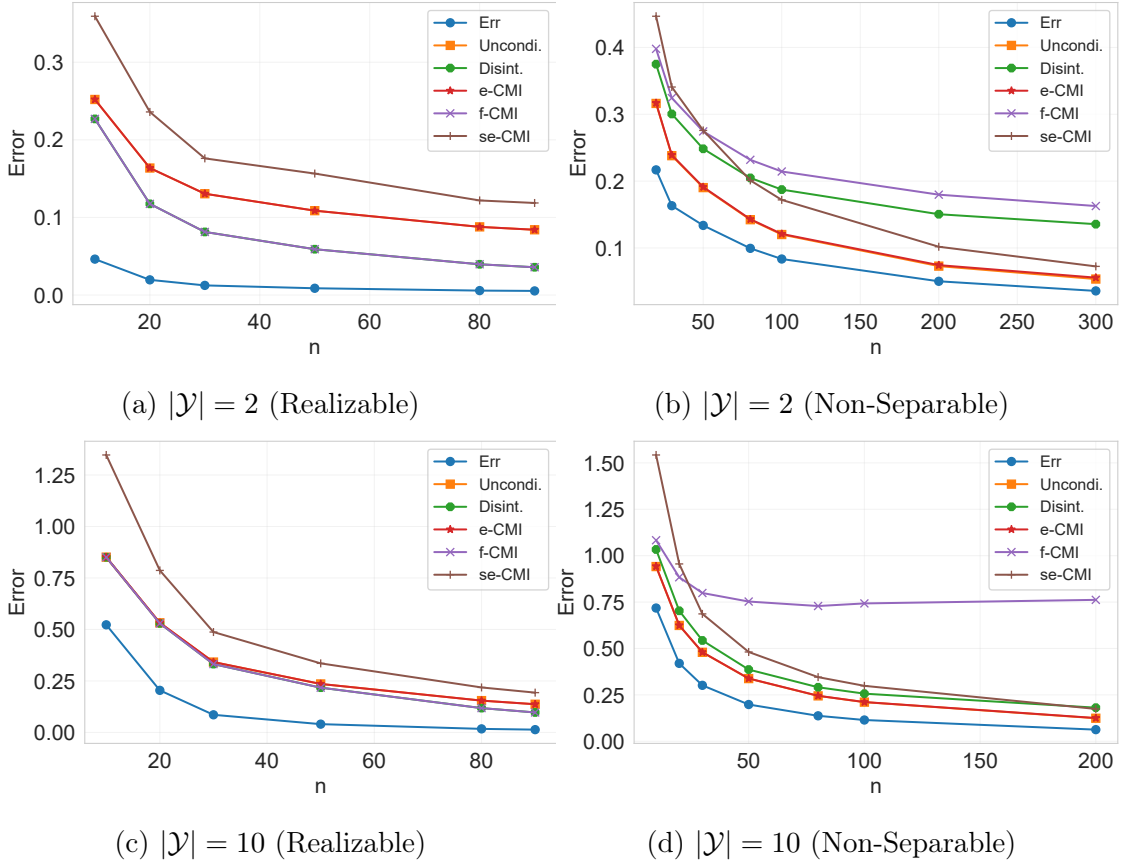


Figure C.4: Comparison of the square-root bounds on the synthetic dataset. Here  $f\text{-CMI}$ ,  $e\text{-CMI}$  and  $se\text{-CMI}$  refer to the disintegrated  $f\text{-CMI}$  bound [82], the unconditional  $e\text{-CMI}$  bound in Corollary and the single-loss based square-root bound in Theorem 5.4.1, respectively.

Estimating the  $\gamma$ -variance and  $\lambda$ -sharpness in the CMI setting is a straightforward process. For example, to estimate sharpness, for each fixed  $\tilde{z}$ , we store the training losses  $L_i^+$  when  $U_i = 0$  (corresponding to  $\tilde{z}^+$ ) and the training losses  $L_i^-$  when  $U_i = 1$  (corresponding

to  $\tilde{z}^-$ ) with different weight configurations  $W$ . By doing so, we collect the necessary data to compute the second term of the equation in Lemma 5.4.5.

### C.5.3 Additional Numerical Results: Lower Bound of the Population Risk Entropy

In Theorem 5.4.2, we give a lower bound for the entropy of the population risk in the interpolating setting, This can be verified experimentally as shown in Figure C.3.

### C.5.4 Additional Numerical Results: Equivalence of Information Stored in Two Columns

In Lemma 5.4.1, we indeed hint that  $P_{L_i^+} = P_{L_i^-}$  and the mutual information  $I(L_i^+; U_i) = I(L_i^-; U_i)$ . Such symmetric argument can also be empirically verified. Figure C.1 further demonstrates that the same amount of evaluated mutual information is stored in the two columns.

### C.5.5 Additional Numerical Results: Interpolating Bounds

We consider the interpolating setting, where Theorem 5.3.3, Eq. (5.3) in Theorem 5.4.3 and the bound of [92, Eq. 12] are all valid. In Figure C.2a, all three bounds are nearly equal to the expected generalization error, while in Figure C.2b, we find that Eq. (5.3) in Theorem 5.4.3 is slightly looser than the other two bounds when  $n \leq 20$ . As mentioned in the main text, the gap between Eq. (5.3) and the bound of [92, Eq. 12] is determined by the difference of  $2I(L_i^+; U_i)$  and  $I(L_i; U_i)$  (i.e. interaction information). From Figure C.2c and Figure C.2d, we further compare these two quantities, and find that  $I(L_i; U_i)$  will be slightly smaller than  $2I(L_i^+; U_i)$  when the number of labels increases and the number of training sample decreases. Notice that [92] indeed uses  $I(L_i; U_i | \tilde{Z})$  instead of  $I(L_i; U_i)$ , which is a looser quantity.

### C.5.6 Additional Numerical Results: Comparison of Square-Root Bounds

We conduct a comparison of square-root bounds on the synthetic dataset, where we also include the disintegrated version of the  $f$ -CMI bound proposed by [82], an improved unconditional e-CMI bound (obtained by replacing  $I(L_i; U_i | \tilde{Z})$  with  $I(L_i; U_i)$ ), and the single-loss

square-root bound presented in Theorem 5.4.1. The results are illustrated in Figure C.4. Consistent with the observations in the main text, we find that the disintegrated bounds are tighter than the unconditional MI bounds when the training loss approaches zero, but looser than the unconditional MI bounds when the training loss is large. This suggests that while, according to the DPI, the unconditional e-CMI bound or ld-MI bound should be tighter than the  $f$ -CMI bound, in some cases, the disintegrated version of the  $f$ -CMI bound may be tighter than the unconditional e-CMI bound or ld-MI bound. For non-separable  $\mu$ , the  $f$ -CMI bound becomes looser as the number of classes increases, which provides justification for the remarks made after Theorem 5.3.1. In fact, it can be even worse than the single-loss square-root bound in Theorem 5.4.1, which includes an undesired constant of 2.

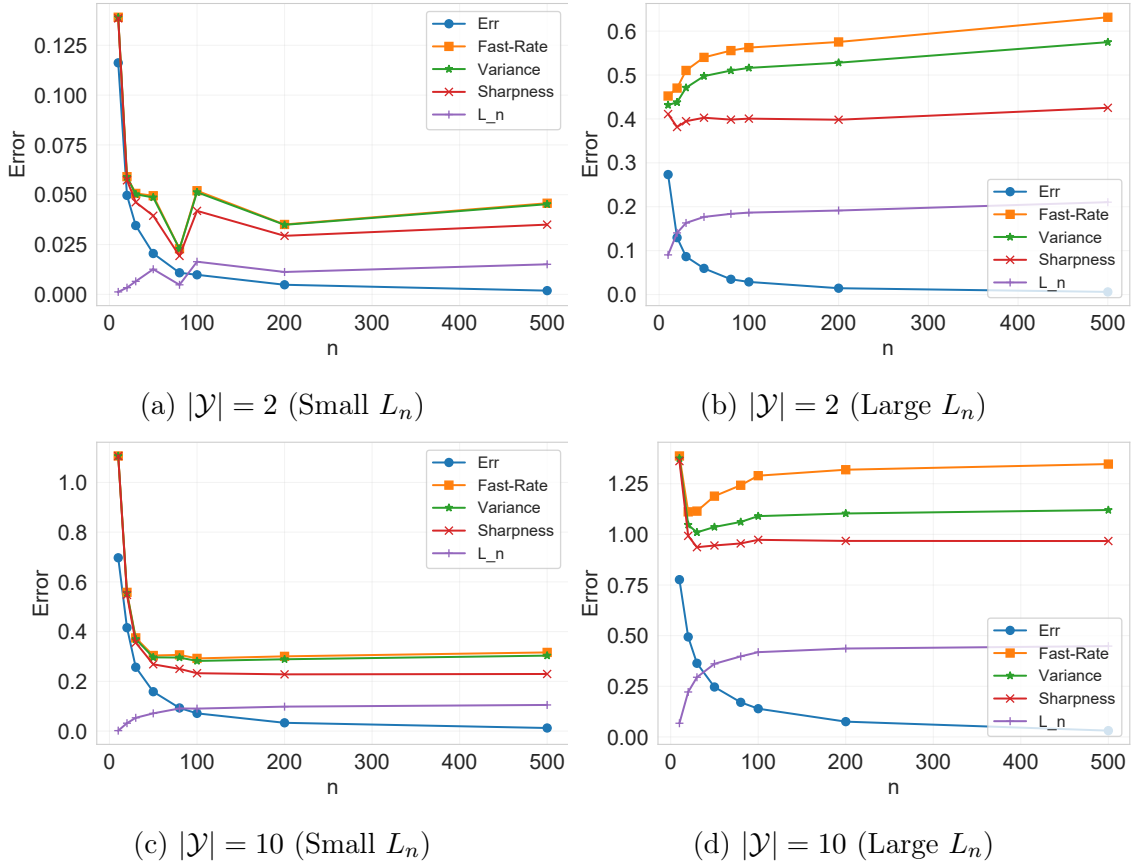


Figure C.5: Comparison of three fast-rate bounds on the synthetic dataset. Here *Fast-Rate* refers to the fast-rate bound of Eq. (5.1) in Theorem 5.4.3.

## C.5.7 Additional Numerical Results: Comparison of Fast-Rate Bounds

We conduct a comparison of fast-rate bounds, including Eq. (5.1) in Theorem 5.4.3, the variance bound in Theorem 5.4.4, and the sharpness bound in Theorem 5.4.5, on the synthetic dataset with fixed values of  $C_1$  and  $C_2$ . As mentioned in the main text, if  $L_n \rightarrow 0$ , both  $V(\gamma)$  and  $F(\lambda)$  become zero, resulting in the three bounds being equivalent. However, when  $L_n \neq 0$ , the variance bound and sharpness bound are always sharper than Eq. (5.1), as discussed earlier. In Figure C.5, we compare these bounds with fixed values of  $C_1 = 3$  and  $C_2 = 0.3$ . Figures C.5a and C.5c demonstrate that when  $L_n$  is small, the gap between the variance bound and Eq. (5.1) is small, indicating that the loss variance in this case is also small. However, the sharpness bound clearly outperforms the other two bounds. Furthermore, in Figures C.5b and C.5d, when  $L_n$  is large, both the sharpness bound and the variance bound significantly improve upon Eq. (5.1). Notably, only the sharpness bound remains non-vacuous in Figure C.5d.

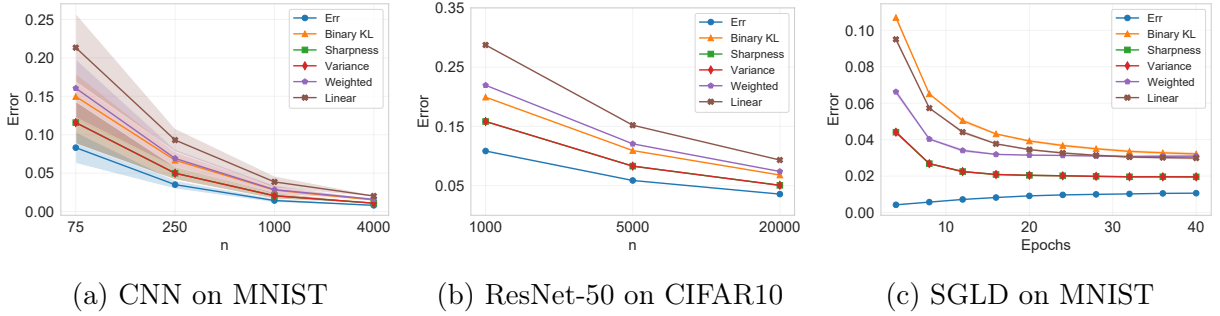


Figure C.6: Comparison of Optimal Fast-Rate Bounds.

# Appendix D

## Appendix of Chapter 6

### D.1 Some Additional Discussions and Clarifications

Similar to [180], in Chapter 6, we mainly focus on 1-Wasserstein distance but all the upper bounds based on 1-Wasserstein distance also holds for higher order Wasserstein distance by Hölder’s inequality [44, Remark 6.6].

Motivated by the classic  $f$ -divergence, [3] proposed a discrepancy measure called  $D_{\mathcal{H}}^{\phi}$ -discrepancy (or  $D_{h,\mathcal{H}}^{\phi}$ -discrepancy). As KL divergence belongs to the family of  $f$ -divergences and both [3] and our work use the variational representation of divergence, there appears to be a connection between our work (in Section 6.4) and theirs. However, it’s important to note that the variational characterization of  $f$ -divergence used in [3] is based on the results of [160], while the Donsker-Varadhan representation of KL divergence (see Lemma 2.4.1) used in our work cannot be directly obtained from their variational characterization [103, 4]. In fact, simply choosing  $x \log x$  as the conjugate function would result in a weaker bound than Lemma 2.4.1. Therefore, while there is some similarity between our results and those of [3], our results in Section 6.4 cannot be directly derived from theirs.

### D.2 Omitted Proofs and Additional Results in Section 6.4

#### D.2.1 Proof of Theorem 6.4.1

*Proof.* Let  $Q = \mu'$ ,  $P = \mu$  and  $g = \ell$ , then Theorem 6.4.1 comes directly from Lemma 2.4.6. □



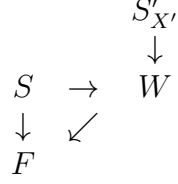


Figure D.1: The relationship between random variables in UDA, where  $F = L_{\mu'}(W) - L_S(W)$ .

### D.2.2 Proof of Corollary 6.4.2

*Proof.* As discussed in Remark 6.4.1, when the loss is bounded in  $[0, M]$ , it is guaranteed to be  $\frac{M}{2}$ -subgaussian for any  $w \in \mathcal{W}$ . Then, similar to the proof of Theorem 6.4.1, let  $Q = \mu$ ,  $P = \mu'$  and  $g = \ell$  and  $R = \frac{M}{2}$ , then the following bound holds by Lemma 2.4.6,

$$|\widetilde{\text{Err}}(w)| \leq \sqrt{\frac{M^2}{2} \text{D}_{\text{KL}}(\mu || \mu')}.$$

Then, by  $\min\{A, B\} \leq \frac{1}{2}(A + B)$ , the remaining part is straightforward,

$$|\widetilde{\text{Err}}(w)| \leq \frac{M}{\sqrt{2}} \sqrt{\min\{\text{D}_{\text{KL}}(\mu || \mu'), \text{D}_{\text{KL}}(\mu' || \mu)\}} \leq \frac{M}{2} \sqrt{\text{D}_{\text{KL}}(\mu || \mu') + \text{D}_{\text{KL}}(\mu' || \mu)}.$$

This completes the proof.  $\square$

### D.2.3 Proof of Theorem 6.4.2

*Proof.* Let  $w^* = \arg \min_{w \in \mathcal{W}} \mathbb{E}_{Z'} [\ell(f_w(X'), Y')] + \mathbb{E}_Z [\ell(f_w(X), Y)]$ . By Lemma 2.4.1,

$$\text{D}_{\text{KL}}(P_{X'} || P_X) \geq \sup_{t \in \mathbb{R}, w \in \mathcal{W}} \mathbb{E}_{X'} [t \ell(f_w(X'), f_{w^*}(X'))] - \log \mathbb{E}_X [e^{t \ell(f_w(X), f_{w^*}(X))}].$$

Recall that  $\ell(f_{w'}(X), f_w(X))$  is  $R$ -subgaussian, by using Lemma 2.4.6 (let  $Q = P_{X'}$ ,  $P = P_X$  and  $g(\cdot) = \ell(f_{w'}(\cdot), f_w(\cdot))$ ), we have

$$|\mathbb{E}_{X'} [\ell(f_w(X'), f_{w^*}(X'))] - \mathbb{E}_X [\ell(f_w(X), f_{w^*}(X))]| \leq \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)}. \quad (\text{D.1})$$

For any  $f_w \in \mathcal{F}$ , by the symmetric and triangle property of the loss, we have

$$\begin{aligned}
& \mathbb{E}_{Z'} [\ell(f_w(X'), Y')] \\
& \leq \mathbb{E}_{X'} [\ell(f_w(X'), f_{w^*}(X'))] + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')] \\
& \leq \mathbb{E}_X [\ell(f_w(X), f_{w^*}(X))] + \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)} + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')] \tag{D.2} \\
& = \int_x \ell(f_w(x), f_{w^*}(x)) dP_X(x) + \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)} + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')] \\
& = \int_x \int_y \ell(f_w(x), f_{w^*}(x)) dP_{Y|X=x}(y) dP_X(x) + \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)} + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')] \\
& \leq \int_x \int_y \ell(f_w(x), y) + \ell(y, f_{w^*}(x)) dP_{Y|X=x}(y) dP_X(x) + \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)} + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')] \tag{D.3} \\
& = \mathbb{E}_Z [\ell(f_w(X), Y)] + \mathbb{E}_Z [\ell(Y, f_{w^*}(X))] + \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)} + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')],
\end{aligned}$$

where Eq. (D.2) is by Eq. (D.1) and Eq. (D.3) is again by the triangle property of the loss function.

Thus,  $\widetilde{\text{Err}}(w) \leq \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)} + \lambda^*$ , which completes the proof.  $\square$

## D.2.4 Proof of Theorem 6.4.3

*Proof.* By Lemma 2.4.1,

$$\begin{aligned}
\text{D}_{\text{KL}}(P_{X'} || P_X) & \geq \sup_{t \in \mathbb{R}, w, w' \in \mathcal{W}^2} \mathbb{E}_{X'} [t\ell(f_w(X'), f_{w'}(X'))] - \log \mathbb{E}_X [e^{t\ell(f_w(X), f_{w'}(X))}] \\
& \geq \sup_{t \in \mathbb{R}} \mathbb{E}_{W, W'} [\mathbb{E}_{X'} [t\ell(f_W(X'), f_{W'}(X'))] - \log \mathbb{E}_X [e^{t\ell(f_W(X), f_{W'}(X))}]] \\
& \geq \sup_{t \in \mathbb{R}} \mathbb{E}_{W, W'} [\mathbb{E}_{X'} [t\ell(f_W(X'), f_{W'}(X'))]] - \log \mathbb{E}_{W, W'} [\mathbb{E}_X [e^{t\ell(f_W(X), f_{W'}(X))}]],
\end{aligned}$$

where the last inequality is by applying Jensen's inequality to the logarithm function, which is concave.

By the subgaussian assumption,

$$|\mathbb{E}_{W, W'} [\mathbb{E}_{X'} [\ell(f_W(X'), f_{W'}(X'))]] - \mathbb{E}_{W, W'} [\mathbb{E}_X [\ell(f_W(X), f_{W'}(X))]]| \leq \sqrt{2R^2 \text{D}_{\text{KL}}(P_{X'} || P_X)}.$$

This concludes the proof.  $\square$

### D.2.5 Proof of Theorem 6.4.4

*Proof.* From the definition, we have

$$\begin{aligned} \left| \widetilde{\text{Err}}(w) \right| &= \left| \mathbb{E}_{Z'} [\ell(f_w(X'), Y')] - \mathbb{E}_Z [\ell(f_w(X), Y)] \right| \\ &\leq \beta \mathbb{W}(\mu, \mu'). \end{aligned}$$

where the last inequality is by the KR duality of Wasserstein distance (see Lemma 2.4.7).  $\square$

### D.2.6 Proof of Corollary 6.4.3

*Proof.* When  $d$  is the discrete metric, Wasserstein distance is equal to the total variation, then by Theorem 6.4.4,

$$\left| \widetilde{\text{Err}}(w) \right| \leq \beta \text{TV}(\mu', \mu),$$

The remaining part is by using Lemma 2.4.8 and Lemma 2.4.9:

$$\beta \text{TV}(\mu', \mu) \leq \beta \sqrt{\min \left\{ \frac{1}{2} \text{D}_{\text{KL}}(\mu' || \mu), 1 - e^{-\text{D}_{\text{KL}}(\mu' || \mu)} \right\}}.$$

Then, if  $\ell$  is bounded by  $M$ , we can replace  $\beta$  by  $M$  above, which completes the proof.  $\square$

### D.2.7 Proof of Theorem 6.4.5

*Proof.* Let  $w^* = \arg \min_{w \in \mathcal{W}} \mathbb{E}_{Z'} [\ell(f_w(X'), Y')] + \mathbb{E}_Z [\ell(f_w(X), Y)]$ .

If  $\ell(f_w(X), f_{w'}(X))$  is  $\beta$ -Lipschitz in  $\mathcal{X}$  for any  $w, w' \in \mathcal{W}$ , then similar to Theorem 6.4.4, it's easy to show that

$$\mathbb{E}_{X'} [\ell(f_w(X'), f^*(X'))] - \mathbb{E}_X [\ell(f_w(X), f^*(X))] \leq \beta \mathbb{W}(P'_X, P_X) \quad (\text{D.4})$$

For any  $f_w \in \mathcal{F}$ , by the symmetric and triangle property of the loss, we have

$$\begin{aligned} &\mathbb{E}_{Z'} [\ell(f_w(X'), Y')] \\ &\leq \mathbb{E}_{X'} [\ell(f_w(X'), f_{w^*}(X'))] + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')] \\ &\leq \mathbb{E}_X [\ell(f_w(X), f_{w^*}(X))] + \beta \mathbb{W}(P'_X, P_X) + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y')] \\ &\leq \mathbb{E}_Z [\ell(f_w(X), Y)] + \mathbb{E}_Z [\ell(Y, f_{w^*}(X))] + \beta \mathbb{W}(P'_X, P_X) + \mathbb{E}_{Z'} [\ell(f_{w^*}(X'), Y)], \end{aligned} \quad (\text{D.5})$$

where Eq. (D.5) is by Eq. (D.4) and the last inequality is again by the triangle property of the loss function. This completes the proof.  $\square$

## D.2.8 Additional Results: Sample Complexity Bounds

One of the main ingredients to derive our sample complexity bound is the following lemma, where a concentration bound for a class of unbounded functions is given.

**Lemma D.2.1** ([40, Corollary 9]). *Let  $\kappa > 2$  and  $\mathcal{G} = \{g : \mathcal{Z} \rightarrow \mathbb{R} \text{ s.t. } \mathbb{E}_\mu [e^{g(Z)}] < +\infty\}$ . Assume  $\mathbb{E}_\mu [g(Z)^\kappa] < +\infty$  for all  $g \in \mathcal{G}$ . Let  $\hat{\mu}$  be the empirical distributions consist of  $n$  data points sampled i.i.d. from  $\mu$ . If  $\mathcal{G}$  has the finite pseudo-dimension  $d$ , then for  $\forall \delta \in (0, 1)$ , the following inequality holds for all  $g \in \mathcal{G}$  with probability at least  $1 - \delta$ ,*

$$\mathbb{E}_\mu [g(Z)] \leq \mathbb{E}_{\hat{\mu}} [g(Z)] + 2\Lambda(\kappa) \sqrt[\kappa]{\mathbb{E}_\mu [g(Z)^\kappa]} \sqrt{\frac{1}{n} \left( d \log \frac{2en}{d} + \log \frac{4}{\delta} \right)},$$

where  $\Lambda(\kappa) = \left(\frac{1}{2}\right)^{\frac{2}{\kappa}} \left(\frac{\kappa}{\kappa-2}\right)^{\frac{\kappa-1}{\kappa}}$ .

Below is another useful lemma for the bounded case, which comes from [145, Theorem 11.8] with a slight modification (by invoking a different VC-dimension based generalization bound from [206]).

**Lemma D.2.2.** *Let  $\mathcal{F} = \{f : \mathcal{Z} \rightarrow \mathbb{R}^+\}$ . Assume  $\mathbb{E}_\mu [f(Z)] < M$  for all  $f \in \mathcal{F}$  for some constant  $M > 0$ . Let  $\hat{\mu}$  be the empirical distributions consist of  $n$  data points sampled i.i.d. from  $\mu$ . If  $\mathcal{F}$  has the finite pseudo-dimension  $d$ , then for  $\forall \delta \in (0, 1)$ , the following inequality holds for all  $f \in \mathcal{F}$  with probability at least  $1 - \delta$ ,*

$$\mathbb{E}_\mu [f(Z)] \leq \mathbb{E}_{\hat{\mu}} [f(Z)] + 2M \sqrt{\frac{1}{n} \left( d \log \frac{2en}{d} + \log \frac{4}{\delta} \right)}.$$

We are now in a position to state our sample complexity bound.

**Theorem D.2.1.** *Let  $\hat{\mu}$  and  $\hat{\mu}'$  be the empirical distributions consist of  $n$  source data and  $m$  target data sampled i.i.d. from  $\mu$  and  $\mu'$ , respectively. Let  $\mathcal{G} = \{g : \mathcal{Z} \rightarrow \mathbb{R} \text{ s.t. } \mathbb{E}_\mu [e^{g(Z)}] < \infty\}$  with finite pseudo-dimension  $d_1$ , and let the pseudo-dimension of  $\{\exp \circ g | g \in \mathcal{G}\}$  be  $d_2$ . Let  $\kappa > 2$  and assume that  $\mathbb{E}_\mu [g(Z)^\kappa] < +\infty$  for all  $g \in \mathcal{G}$ . Assume there exists a*

constant  $\alpha \leq \min_{g \in \mathcal{G}} \{\mathbb{E}_{\hat{\mu}} [e^{g(Z)}], \mathbb{E}_{\mu} [e^{g(Z)}]\}$ . Then for  $\forall \delta \in (0, 1)$  the following bound holds with probability at least  $1 - \delta$ ,

$$D_{\text{KL}}(\mu' || \mu) - D_{\text{KL}}(\hat{\mu}' || \hat{\mu}) \leq C_1(\kappa) \sqrt{\frac{1}{n} \left( d_1 \log \frac{2en}{d_1} + \log \frac{4}{\delta} \right)} + C_2(\alpha) \sqrt{\frac{1}{m} \left( d_2 \log \frac{2em}{d_2} + \log \frac{4}{\delta} \right)},$$

where  $C_1(\kappa) = \left(\frac{1}{2}\right)^{\frac{2-\kappa}{\kappa}} \left(\frac{\kappa}{\kappa-2}\right)^{\frac{\kappa-1}{\kappa}} \sup_{g \in \mathcal{G}} \sqrt{\kappa \mathbb{E}_{\mu} [g(Z)^{\kappa}]}$  and  $C_2(\alpha) = \frac{2}{\alpha} \sup_{g \in \mathcal{G}} \mathbb{E}_{\mu} [e^{g(Z)}]$ .

*Proof.* Recall Lemma 2.4.1, we have

$$D_{\text{KL}}(\mu' || \mu) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'} [g(Z')] - \log \mathbb{E}_{\mu} [e^{g(Z)}],$$

and

$$D_{\text{KL}}(\hat{\mu}' || \hat{\mu}) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\hat{\mu}'} [g(Z')] - \log \mathbb{E}_{\hat{\mu}} [e^{g(Z)}].$$

Then, with the probability at least  $1 - \delta$ ,

$$\begin{aligned} & D_{\text{KL}}(\mu' || \mu) - D_{\text{KL}}(\hat{\mu}' || \hat{\mu}) \\ &= \sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'} [g(Z')] - \log \mathbb{E}_{\mu} [e^{g(Z)}] - \left( \sup_{g \in \mathcal{G}} \mathbb{E}_{\hat{\mu}'} [g(Z')] - \log \mathbb{E}_{\hat{\mu}} [e^{g(Z)}] \right) \\ &\leq \sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'} [g(Z')] - \log \mathbb{E}_{\mu} [e^{g(Z)}] - (\mathbb{E}_{\hat{\mu}'} [g(Z')] - \log \mathbb{E}_{\hat{\mu}} [e^{g(Z)}]) \\ &= \sup_{g \in \mathcal{G}} \mathbb{E}_{\mu'} [g(Z')] - \mathbb{E}_{\hat{\mu}'} [g(Z')] + \log \mathbb{E}_{\hat{\mu}} [e^{g(Z)}] - \log \mathbb{E}_{\mu} [e^{g(Z)}] \\ &\leq \sup_{g \in \mathcal{G}} |\mathbb{E}_{\mu'} [g(Z')] - \mathbb{E}_{\hat{\mu}'} [g(Z')]| + \sup_{g \in \mathcal{G}} |\log \mathbb{E}_{\hat{\mu}} [e^{g(Z)}] - \log \mathbb{E}_{\mu} [e^{g(Z)}]| \\ &\leq \sup_{g \in \mathcal{G}} |\mathbb{E}_{\mu'} [g(Z')] - \mathbb{E}_{\hat{\mu}'} [g(Z')]| + \sup_{g \in \mathcal{G}} \frac{1}{\alpha} |\mathbb{E}_{\hat{\mu}} [e^{g(Z)}] - \mathbb{E}_{\mu} [e^{g(Z)}]| \end{aligned} \quad (\text{D.6})$$

$$\leq C_1(\kappa) \sqrt{\frac{1}{n} \left( d_1 \log \frac{2en}{d_1} + \log \frac{4}{\delta} \right)} + C_2(\alpha) \sqrt{\frac{1}{m} \left( d_2 \log \frac{2em}{d_2} + \log \frac{4}{\delta} \right)}, \quad (\text{D.7})$$

where Eq. (D.6) is derived below.

W.L.O.G. assume that  $\mathbb{E}_{\hat{\mu}} [e^{g(Z)}] \leq \mathbb{E}_{\mu} [e^{g(Z)}]$  (and Eq. (D.6) still holds when  $\mathbb{E}_{\hat{\mu}} [e^{g(Z)}] \geq$

$\mathbb{E}_\mu [e^{g(Z)}]$ ), then

$$\begin{aligned}
\left| \log \mathbb{E}_{\hat{\mu}} [e^{g(Z)}] - \log \mathbb{E}_\mu [e^{g(Z)}] \right| &= \left| \log \frac{\mathbb{E}_\mu [e^{g(Z)}]}{\mathbb{E}_{\hat{\mu}} [e^{g(Z)}]} \right| = \left| \log \left( 1 + \frac{\mathbb{E}_\mu [e^{g(Z)}]}{\mathbb{E}_{\hat{\mu}} [e^{g(Z)}]} - 1 \right) \right| \\
&\leq \left| \frac{\mathbb{E}_\mu [e^{g(Z)}]}{\mathbb{E}_{\hat{\mu}} [e^{g(Z)}]} - 1 \right| \\
&= \left| \frac{1}{\mathbb{E}_{\hat{\mu}} [e^{g(Z)}]} (\mathbb{E}_\mu [e^{g(Z)}] - \mathbb{E}_{\hat{\mu}} [e^{g(Z)}]) \right| \\
&\leq \frac{1}{\alpha} |\mathbb{E}_\mu [e^{g(Z)}] - \mathbb{E}_{\hat{\mu}} [e^{g(Z)}]|.
\end{aligned}$$

Eq. (D.7) is by Lemma D.2.1 and Lemma D.2.2. This concludes the proof.  $\square$

With Theorem D.2.1 and Theorem 6.4.1, we immediately have the following corollary.

**Corollary D.2.1.** *Let the conditions in Theorem D.2.1 and Theorem 6.4.1 hold, then for any  $w \in \mathcal{W}$ ,*

$$\left| \widetilde{\text{Err}}(w) \right| \leq \sqrt{2}R \sqrt{D_{\text{KL}}(\hat{\mu}' || \hat{\mu}) + C_1(\kappa) \sqrt{\frac{1}{n} \left( d_1 \log \frac{2en}{d_1} + \log \frac{4}{\delta} \right)} + C_2(\alpha) \sqrt{\frac{1}{m} \left( d_2 \log \frac{2em}{d_2} + \log \frac{4}{\delta} \right)}},$$

where  $C_1(\kappa)$  and  $C_2(\alpha)$  are the same as in Theorem D.2.1.

## D.2.9 Additional Discussions on the Convergence of Empirical KL Divergence

Characterizing the convergence of the empirical KL divergence to the real KL is a challenging task that often requires several additional assumptions, as demonstrated in Theorem D.2.1. However, it is worth noting that the convergence rate of the empirical distribution to the real distribution in the KL sense is already established in the discrete space. This fact is supported by a classic result in [41, Theorem 11.2.1], which we state in the following theorem:

**Theorem D.2.2.** *Let  $\hat{\mu}$  and  $\hat{\mu}'$  be defined as in Theorem D.2.1. Assume the space of  $\mathcal{Z}$  is finite (i.e.  $|\mathcal{Z}| \leq \infty$ ), then for  $\forall \delta \in (0, 1)$ , with probability at least  $1 - \delta$ ,*

$$D_{\text{KL}}(\hat{\mu} || \mu) \leq \frac{|\mathcal{Z}|}{n} \log(n+1) + \frac{1}{n} \log \frac{1}{\delta}, \quad D_{\text{KL}}(\hat{\mu}' || \mu') \leq \frac{|\mathcal{Z}|}{m} \log(m+1) + \frac{1}{m} \log \frac{1}{\delta}.$$

Thus, it suffices to ensure that the empirical KL converge to the real KL with the similar rate, although we do not know if there might exist a faster convergence rate.

### D.2.10 Generalize to Approximate Triangle Inequality

In Section 6.4, some results require that the loss obeys the triangle inequality (i.e. Assumption 6.4), such as Theorem 6.4.2 and Theorem 6.4.5. While the 0 – 1 loss satisfies Assumption 6.4, some other loss may not. Thus, to generalize Theorem 6.4.2 and Theorem 6.4.5, we invoke an approximate triangle inequality, which is originally defined in [42].

**Assumption D.1** ( $\alpha$ -Triangle).  $\ell(\cdot, \cdot)$  is symmetric and satisfies the following  $\alpha$ -triangle inequality:  $\ell(y_1, y_2) \leq \alpha(\ell(y_1, y_3) + \ell(y_3, y_2))$  for any  $y_1, y_2, y_3 \in \mathcal{Y}$ , where  $\alpha \geq 1$  is a constant that may depend on the hypothesis space  $\mathcal{W}$  and the loss  $\ell$ .

**Remark D.2.1.** We note that the squared loss satisfies 2-triangle inequality.

Thus, Theorem 6.4.2 can be easily generalized below.

**Theorem D.2.3.** If Assumption D.1 holds and let  $\ell(f_{w'}(X), f_w(X))$  be  $R$ -subgaussian for any  $w, w' \in \mathcal{W}$ . Then for any  $w$ ,

$$\widetilde{\text{Err}}(w) \leq (\alpha^2 - 1)L_\mu + \alpha\sqrt{2R^2\text{D}_{\text{KL}}(P_{X'}||P_X)} + \alpha^2\lambda^*,$$

where  $\lambda^* = \min_{w \in \mathcal{W}} L_{\mu'}(w) + L_\mu(w)$ .

Theorem 6.4.5 can be generalized in the similar way. While Theorem D.2.3 strictly speaking is not a generalization bound, as it includes  $L_\mu$  in the bound, it shares the same underlying concept as Theorem 6.4.2. Namely, to minimize the population risk in the target domain, it is essential for the source domain and target domain to be similar, and for both  $L_\mu$  and  $\lambda^*$  to be kept small.

## D.3 Omitted Proofs and Additional Discussions in Section 6.5

### D.3.1 Additional Discussion on Theorem 6.5.1

To derive the bound in Theorem 6.5.1, we need to make use of the second equality in Eq. (6.1). In fact, by the definition of  $\text{Err}$  (the first equality in Eq. (6.1)), the unlabelled sample

$S'_{X'_j}$  does not explicitly appear, so one can easily apply the similar information-theoretic analysis starting from the first equality in Eq. (6.1), and obtain an upper bound that consists of  $I(W; Z_i)$  and  $D_{\text{KL}}(\mu||\mu')$ . Precisely, the following bound holds,

**Theorem D.3.1.** *Assume  $\ell(f_w(X'), Y')$  is  $R$ -subgaussian for any  $w \in \mathcal{W}$ . Then*

$$|\text{Err}| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \sqrt{2R^2 I(W; Z_i)} + \sqrt{2R^2 D_{\text{KL}}(\mu||\mu')}.$$

The proof of Theorem D.3.1 is nearly the same to the proof of [225, Corollary 2] and [139, Corollary 1].

It's important to note that although

$$I(W; Z_i) \leq I(W; Z_i | X'_j) = \mathbb{E}_{X'_j} \left[ I^{X'_j}(W; Z_i) \right],$$

the bound in Theorem 6.5.1 is incomparable to the bound based on  $I(W; Z_i)$ . This is mainly due to the fact that we use the disintegrated version of mutual information,  $I^{X'_j}(W; Z_i)$ , and the expectation over  $X'_j$  is outside of the square root, which is a convex function. Using  $I^{X'_j}(W; Z_i)$  instead of  $I(W; Z_i)$  allows us to figure out more details about the role of unlabelled target data in the algorithm. Additionally, one can also prove a bound based on  $I(W; Z_i | X'_j)$  (e.g., simply applying Jensen's inequality to Theorem 6.5.1), which is close to an individual and UDA version of [28, Theorem 3].

Furthermore, the first term in Theorem 6.5.1 characterize the expected generalization gap on the source domain (i.e.  $\mathbb{E}_{W,S} [L_\mu(W) - L_S(W)]$ ), then the bound suggests us that it's possible to invoke the unlabelled target data to further improve the performance on source domain, and the simplest case is the semi-supervised learning (when  $\mu = \mu'$ ).

**Compared with [225, 105].** Notably, bounds in [225, 105] fail to characterize the dependence between  $W$  and  $S'_{X'}$ . More precisely, the algorithm-dependent term in their bounds is  $I(W; Z_i)$  or  $I(W; S)$ , while our algorithm-dependent term is  $I^{X'_j}(W; Z_i)$  that directly depends on the unlabelled target data. Moreover, while the disintegrated mutual information  $I^{X'_j}(W; Z_i)$  and the unconditional mutual information  $I(W; Z_i)$  cannot be directly compared, recent work by [217] provides empirical evidence comparing similar terms in the supervised learning setting. Specifically, they demonstrate that when the empirical risk is small, such as in a realizable case, the disintegrated mutual information is smaller than the unconditional mutual information. Conversely, when the empirical risk is large, the unconditional mutual information is the smaller of the two.



**More Discussion on the Vanishing of  $I(X'_j; Z_i|W)$  in Remark 6.5.1.** Note that  $S$  depends on  $S'_{X'}$  given  $W$ , so intuitively the dependence between each individual instance  $Z_i$  and  $X'_j$  is weaker when  $n$  and  $m$  become larger. More precisely, W.L.O.G let  $i = j = 1$ , and recall that  $W = \mathcal{A}(S'_{X'}, S)$ , when  $n, m \rightarrow \infty$ , taking  $S$  and  $S'_{X'}$  as the input of the algorithm is nearly equivalent to computing  $W$  based on the source distribution  $\mu$  and the target distribution  $P_{X'}$ , thus,  $W$  will only depend on the two distributions, without depending on the realizations  $Z_1$  and  $X'_1$  drawn respectively from the two distributions, that is,  $I(Z_1; X'_1|W) = I(Z_1; X'_1|\mathcal{A}(\mu, P_{X'})) = I(Z_1; X'_1) = 0$ . In addition, one may argue that what if  $W = \text{constant}$  that does not really depend on the input data. In this case,  $I(Z_1; X'_1|W) = I(Z_1; X'_1) = 0$  will hold trivially. In the other extreme, if  $n = 1$  and  $m = 1$ , then  $W = \mathcal{A}(X'_1, Z_1)$ , and the quantity  $I(Z_1; X'_1|\mathcal{A}(X'_1, Z_1))$  should be large. When  $n$  and  $m$  increase, it becomes  $I(Z_1; X'_1|\mathcal{A}(X'_{1:m}, Z_{1:n}))$ . Now we want to guess  $Z_1$  from  $X'_1$ , this should be easier when having the knowledge of  $\mathcal{A}(X'_1; Z_1)$  compared with when having the knowledge of  $\mathcal{A}(X'_{1:m}; Z_{1:n})$ .

### D.3.2 Proof of Theorem 6.5.1

*Proof.* By Lemma 2.4.1,

$$\begin{aligned}
& \text{D}_{\text{KL}} \left( P_{W, Z_i | X'_j = x'_j} || P_{W, Z' | X'_j = x'_j} \right) \\
&= \text{D}_{\text{KL}} \left( P_{W, Z_i | X'_j = x'_j} || P_{W | X'_j = x'_j} P_{Z'} \right) \\
&\geq \sup_t \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} [t\ell(f_W(X_i), Y_i)] - \log \mathbb{E}_{P_{W | X'_j = x'_j} P_{Z'}} [\exp(t\ell(f_W(X'), Y'))] \\
&\geq \sup_t \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} [t\ell(f_W(X_i), Y_i)] - \mathbb{E}_{P_{W | X'_j = x'_j}} [tL_{\mu'}(W)] - R^2 t^2 / 2,
\end{aligned} \tag{D.8}$$

where Eq. (D.8) is by the independence between algorithm output  $W$  and unseen target domain data  $Z'$ , and the last inequality is by the subgaussian assumption.

Thus,

$$\left| \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{P_{W | X'_j = x'_j}} [L_{\mu'}(W)] \right| \leq \sqrt{2R^2 \text{D}_{\text{KL}} \left( P_{W, Z_i | X'_j = x'_j} || P_{W | X'_j = x'_j} P_{Z'} \right)}. \tag{D.9}$$

Exploiting the fact that

$$\begin{aligned}
|\text{Err}| &= \left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W,Z_i} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W,Z'} [\ell(f_W(X'), Y')] \right| \\
&= \left| \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{X'_j} \left[ \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W,Z_i|X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W,Z'|X'_j} [\ell(f_W(X'), Y')] \right] \right| \\
&\leq \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{X'_j} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W,Z_i|X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W,Z'|X'_j} [\ell(f_W(X'), Y')] \right| \\
&\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \left| \mathbb{E}_{W,Z_i|X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W|X'_j} [L_{\mu'}(W)] \right|,
\end{aligned}$$

where the last two inequalities are by the Jensen's inequality for the absolute function.

Notice that

$$\begin{aligned}
D_{\text{KL}} \left( P_{W,Z_i|X'_j=x'_j} || P_{W|X'_j=x'_j} P_{Z'} \right) &= \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}} \left[ \log \frac{P_{W,Z_i|X'_j=x'_j}}{P_{W|X'_j=x'_j} P_{Z'}} \right] \\
&= \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}} \left[ \log \frac{P_{W|Z_i,X'_j=x'_j} P_{Z_i}}{P_{W|X'_j=x'_j} P_{Z'}} \right] \\
&= \mathbb{E}_{P_{W,Z_i|X'_j=x'_j}} \left[ \log \frac{P_{W|Z_i,X'_j=x'_j}}{P_{W|X'_j=x'_j}} \right] + \mathbb{E}_{P_{Z_i}} \left[ \log \frac{P_{Z_i}}{P_{Z'}} \right] \\
&= I(W; Z_i | X'_j = x'_j) + D_{\text{KL}}(\mu || \mu').
\end{aligned}$$

Recall Eq. (D.9), we then have

$$\begin{aligned}
|\text{Err}| &\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \left| \mathbb{E}_{W,Z_i|X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W|X'_j} [L_{\mu'}(W)] \right| \\
&\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{2R^2 D_{\text{KL}} \left( P_{W,Z_i|X'_j} || P_{W|X'_j} P_{Z'} \right)} \\
&= \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{2R^2 (I^{X'_j}(W; Z_i) + D_{\text{KL}}(\mu || \mu'))} \\
&\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{2R^2 I^{X'_j}(W; Z_i)} + \sqrt{2R^2 D_{\text{KL}}(\mu || \mu')}.
\end{aligned}$$

This completes the proof. □

### D.3.3 Proof of Corollary 6.5.1

*Proof.* We now modify the proof in Theorem 6.5.1.

Recall that

$$|\text{Err}| \leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \left| \mathbb{E}_{W, Z_i | X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j} [L_{\mu'}(W)] \right|.$$

We first decompose the right hand side,

$$\begin{aligned} & \left| \mathbb{E}_{W, Z_i | X'_j = x'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j = x'_j} [L_{\mu'}(W)] \right| \\ = & \left| \mathbb{E}_{W, Z_i | X'_j = x'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j = x'_j} [L_{\mu}(W)] + \mathbb{E}_{W | X'_j = x'_j} [L_{\mu}(W)] - \mathbb{E}_{W | X'_j = x'_j} [L_{\mu'}(W)] \right| \\ \leq & \left| \mathbb{E}_{W, Z_i | X'_j = x'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j = x'_j} [L_{\mu}(W)] \right| + \left| \mathbb{E}_{W | X'_j = x'_j} [L_{\mu}(W) - L_{\mu'}(W)] \right| \\ \leq & \left| \mathbb{E}_{W, Z_i | X'_j = x'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j = x'_j} [L_{\mu}(W)] \right| + \frac{M}{\sqrt{2}} \sqrt{\min\{\text{D}_{\text{KL}}(\mu || \mu'), \text{D}_{\text{KL}}(\mu' || \mu)\}}, \end{aligned}$$

where the last inequality is by Corollary 6.4.2.

Then for the first term in RHS, notice that

$$\begin{aligned} & \text{D}_{\text{KL}} \left( P_{W, Z | X'_j = x'_j} || P_{W, Z_i | X'_j = x'_j} \right) \\ = & \text{D}_{\text{KL}} \left( P_{W | X'_j = x'_j} P_Z || P_{W, Z_i | X'_j = x'_j} \right) \\ \geq & \sup_t \mathbb{E}_{P_{W | X'_j = x'_j} P_Z} [t\ell(f_W(X), Y)] - \log \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} [\exp t\ell(f_W(X_i), Y_i)] \\ \geq & \sup_t \mathbb{E}_{P_{W | X'_j = x'_j} P_Z} [t\ell(f_W(X), Y)] - \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} [t\ell(f_W(X_i), Y_i)] \\ & - \log \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} \left[ e^{t(\ell(f_W(X_i), Y_i) - \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} [\ell(f_W(X_i), Y_i)])} \right] \\ \geq & \sup_t \mathbb{E}_{P_{W | X'_j = x'_j}} [tL_{\mu}(W)] - \mathbb{E}_{P_{W, Z_i | X'_j = x'_j}} [t\ell(f_W(X_i), Y_i)] - M^2 t^2 / 8, \end{aligned}$$

where the last inequality is due to the fact that  $\ell$  is bounded by  $M$  and  $\ell(f_W(X_i), Y_i)$  is  $M/2$ -subgaussian.

Thus,

$$\begin{aligned} \left| \mathbb{E}_{W, Z_i | X'_j = x'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j = x'_j} [L_\mu(W)] \right| &\leq \sqrt{\frac{M^2}{2} \text{D}_{\text{KL}} \left( P_{W | X'_j = x'_j} P_Z || P_{W, Z_i | X'_j = x'_j} \right)} \\ &= \sqrt{\frac{M^2}{2} L(W, Z_i | X'_j = x'_j)}. \end{aligned}$$

Plugging this inequality with the decomposition into the inequality at the beginning of the proof, we have

$$|\text{Err}| \leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{\frac{M^2}{2} L^{X'_j}(W, Z_i)} + \frac{M}{\sqrt{2}} \sqrt{\min\{\text{D}_{\text{KL}}(\mu || \mu'), \text{D}_{\text{KL}}(\mu' || \mu)\}}.$$

Similar development also holds for  $\text{D}_{\text{KL}} \left( P_{W, Z_i | X'_j = x'_j} || P_{W | X'_j = x'_j} P_Z \right)$  as in the proof of Theorem 6.5.1, thus

$$|\text{Err}| \leq \frac{M}{\sqrt{2nm}} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{\min \left\{ I^{X'_j}(W; Z_i), L^{X'_j}(W; Z_i) \right\}} + \frac{M}{\sqrt{2}} \sqrt{\min \{ \text{D}_{\text{KL}}(\mu || \mu'), \text{D}_{\text{KL}}(\mu' || \mu) \}}.$$

This completes the proof.  $\square$

### D.3.4 Proof of Theorem 6.5.2

*Proof.* Similar to the proof of Corollary 6.5.1, recall Theorem 6.4.4,

$$\begin{aligned} &|\text{Err}| \\ &\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \left| \mathbb{E}_{W, Z_i | X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j} [L_\mu(W)] + \mathbb{E}_{W | X'_j} [L_\mu(W)] - \mathbb{E}_{W | X'_j} [L_{\mu'}(W)] \right| \\ &\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \left| \mathbb{E}_{W, Z_i | X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j} [L_\mu(W)] \right| + \beta \mathbb{W}(\mu, \mu') \\ &\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j, Z_i} \left| \mathbb{E}_{W | Z_i, X'_j} [\ell(f_W(X_i), Y_i)] - \mathbb{E}_{W | X'_j} [\ell(f_W(X_i), Y_i)] \right| + \beta \mathbb{W}(\mu, \mu') \\ &\leq \frac{\beta'}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j, Z_i} \mathbb{W}(P_{W | X'_j, Z_i}, P_{W | X'_j}) + \beta \mathbb{W}(\mu, \mu'), \end{aligned}$$

where the last inequality is by Lemma 2.4.7. This concludes the proof.  $\square$

### D.3.5 Proof of Corollary 6.5.2

*Proof.* Similar to the proof of Corollary 6.4.3, replacing Wasserstein distance by the total variation and replacing  $\beta$  and  $\beta'$  by  $M$ , will give us the first inequality,

$$|\widetilde{\text{Err}}| \leq \frac{M}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j, Z_i} \left[ \text{TV}(P_{W|Z_i, X'_j}, P_{W|X'_j}) \right] + M \text{TV}(\mu, \mu').$$

The second inequality is by Lemma 2.4.8,

$$|\widetilde{\text{Err}}| \leq \frac{M}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j, Z_i} \sqrt{\frac{1}{2} \text{D}_{\text{KL}}(P_{W|Z_i, X'_j} \| P_{W|X'_j})} + \sqrt{\frac{M^2}{2} \text{D}_{\text{KL}}(\mu \| \mu')}.$$

Again, one can also apply Lemma 2.4.9 here. This concludes the proof.  $\square$

### D.3.6 Proof of Theorem 6.5.3

*Proof.* Recall Theorem 6.5.1 and by Jensen's inequality we have

$$\begin{aligned} |\text{Err}| &\leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n \mathbb{E}_{X'_j} \sqrt{2R^2 I^{X'_j}(W; Z_i) + \sqrt{2R^2 \text{D}_{\text{KL}}(\mu \| \mu')}} \\ &\leq \sqrt{\frac{2R^2}{nm} \sum_{j=1}^m \sum_{i=1}^n I(W; Z_i | X'_j) + \sqrt{2R^2 \text{D}_{\text{KL}}(\mu \| \mu')}}. \end{aligned}$$

Let  $X'_{1, \dots, j-1, j+1, \dots, m} = S'_{X'} \setminus X'_j$ . Notice that

$$\begin{aligned} I(W; Z_i | S'_{X'}) &= I(W; Z_i | S'_{X'}) + I(X'_{1, \dots, j-1, j+1, \dots, m}; Z_i | X'_j) \\ &= I(W; Z_i | X'_j) + I(X'_{1, \dots, j-1, j+1, \dots, m}; Z_i | X'_j, W) \\ &\geq I(W; Z_i | X'_j). \end{aligned}$$

Thus,  $I(W; Z_i | X'_j) \leq I(W; Z_i | S'_{X'})$ . Then

$$\frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n I(W; Z_i | X'_j) \leq \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n I(W; Z_i | S'_{X'}) = \frac{1}{n} \sum_{i=1}^n I(W; Z_i | S'_{X'}).$$

Then, since  $S \perp\!\!\!\perp S'_{X'}$  and  $Z_i \perp\!\!\!\perp Z_{1:i-1}$  for any  $i \in [n]$ , by the chain rule of mutual information, we have

$$\begin{aligned}
I(W; S|S'_{X'}) &= \sum_{i=1}^n I(W; Z_i|S'_{X'}, Z_{1:i-1}) = \sum_{i=1}^n I(W; Z_i|S'_{X'}, Z_{1:i-1}) + I(Z_i; Z_{1:i-1}) \\
&= \sum_{i=1}^n I(W, Z_{1:i-1}; Z_i|S'_{X'}) \\
&= \sum_{i=1}^n I(W; Z_i|S'_{X'}) + I(Z_i; Z_{1:i-1}|S'_{X'}, W) \\
&\geq \sum_{i=1}^n I(W; Z_i|S'_{X'}).
\end{aligned}$$

Thus, the generalization error bound becomes

$$|\text{Err}| \leq \sqrt{\frac{2R^2}{n} I(W; S|S'_{X'})} + \sqrt{2R^2 \text{D}_{\text{KL}}(\mu||\mu')}.$$

Recall the updating rule of  $W$  and notice that  $W_0$  is independent of  $S$  and  $S'_{X'}$ , the following process is by using the chain rule of mutual information and data processing inequality recurrently,

$$\begin{aligned}
I(W_T; S|S'_{X'}) &= I(W_{T-1} - \eta_T g(W_{T-1}, Z_{B_T}, X'_{B_T}) + N_T; S|S'_{X'}) \\
&\leq I(W_{T-1}, -\eta_T g(W_{T-1}, Z_{B_T}, X'_{B_T}) + N_T; S|S'_{X'}) \\
&= I(W_{T-1}; S|S'_{X'}) + I(\eta_T g(W_{T-1}, Z_{B_T}, X'_{B_T}) + N_T; S|S'_{X'}, W_{T-1}) \\
&\vdots \\
&= \sum_{t=1}^T I(\eta_t g(W_{t-1}, Z_{B_t}, X'_{B_t}) + N_t; S|S'_{X'}, W_{t-1}).
\end{aligned}$$

For each  $t \in [T]$ , denote  $g(W_{t-1}, Z_{B_t}, X'_{B_t})$  as  $G_t$ , then

$$\begin{aligned}
I(\eta_t g(W_{t-1}, Z_{B_t}, X'_{B_t}) + N_t; S|S'_{X'}, W_{t-1}) &= \mathbb{E}_{S'_{X'}, W_{t-1}, S} \left[ \text{D}_{\text{KL}}(P_{G_t + \frac{N_t}{\eta_t} | S, S'_{X'}, W_{t-1}} || P_{G_t + \frac{N_t}{\eta_t} | S'_{X'}, W_{t-1}}) \right] \\
&\leq \mathbb{E}_{S'_{X'}, W_{t-1}, S} \left[ \text{D}_{\text{KL}}(P_{G_t + \frac{N_t}{\eta_t} | S, S'_{X'}, W_{t-1}} || P_{\mathbb{E}_S[G_t] + \frac{N_t}{\eta_t} | S'_{X'}, W_{t-1}}) \right] \\
&= \frac{\eta_t^2}{2\sigma_t^2} \mathbb{E}_{S'_{X'}, W_{t-1}, S} [||G_t - \mathbb{E}_S[G_t]||^2],
\end{aligned}$$

where the inequality is by Lemma 2.4.10 and the last equality is by the KL divergence between two Gaussian distributions.

Finally, putting everything together,

$$|\text{Err}| \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \frac{\eta_t^2}{\sigma_t^2} \mathbb{E}_{S'_{X'}, W_{t-1}, S} [\|G_t - \mathbb{E}_S[G_t]\|^2]} + \sqrt{2R^2 \text{D}_{\text{KL}}(\mu||\mu')},$$

which concludes the proof.  $\square$

### D.3.7 Derivation of Eq. (6.6)

Recall the expected cross-entropy loss, we have

$$\begin{aligned} \mathbb{E}_{W, Z_i} [\ell(f_W(T_i), Y_i)] &= \mathbb{E}_{Z_i, W} [-\log Q_{Y_i|T_i, W}] \\ &= \mathbb{E}_{Z_i, W} \left[ \log \frac{P_{Y_i|T_i, W}}{Q_{Y_i|T_i, W} P_{Y_i|T_i, W}} \right] \\ &= H(Y_i|T_i, W) + \mathbb{E}_{X_i, W} [\text{D}_{\text{KL}}(P_{Y_i|T_i, W} || Q_{Y_i|T_i, W})] \\ &= \mathbb{E}_{Z_i, W} \left[ \log \frac{P_{Y_i|T_i} P_{W|T_i}}{P_{Y_i|T_i, W} P_{Y_i|T_i} P_{W|T_i}} \right] + \mathbb{E}_{T_i, W} [\text{D}_{\text{KL}}(P_{Y_i|T_i, W} || Q_{Y_i|T_i, W})] \\ &= \mathbb{E}_{Z_i, W} \left[ \log \frac{P_{Y_i|T_i} P_{W|T_i}}{P_{Y_i, W|T_i} P_{Y_i|T_i}} \right] + \mathbb{E}_{T_i, W} [\text{D}_{\text{KL}}(P_{Y_i|T_i, W} || Q_{Y_i|T_i, W})] \\ &= H(Y_i|T_i) - I(W; Y_i|T_i) + \mathbb{E}_{T_i, W} [\text{D}_{\text{KL}}(P_{Y_i|T_i, W} || Q_{Y_i|T_i, W})]. \end{aligned}$$

### D.3.8 Additional Discussion on LIMIT

In Section 6.5, we discussed the LIMIT approach proposed by [83] as a means of controlling label information memorization during training. Roughly speaking, to update the classifier parameters, LIMIT constructs an auxiliary network that predicts gradients instead of using the true gradients, which avoids direct use of the true labels for training. To obtain accurate gradients, the auxiliary network needs to be trained using the true labels. We found that the training of LIMIT is unstable and difficult to tune the hyperparameters when used under UDA settings. Therefore, we opted to use the pseudo label strategy proposed in Section 6.5 instead of the pseudo gradient strategy.

## D.4 Experiment Details

We implemented our approach using PyTorch [164] and conducted all experiments on NVIDIA Tesla V100 GPUs with 32 GB of memory. Our code builds largely on the implementation from [70]<sup>1</sup> and [158]<sup>2</sup>.

### D.4.1 Objective Functions of Gradient Penalty and Controlling Label Information

For every iteration, the objective function after adding the gradient penalty becomes

$$\min_W \hat{L}(W, Z_{B_t}, X'_{B_t}) + \lambda_1 \|g(W, Z_{B_t}, X'_{B_t})\|^2,$$

where  $\hat{L}(W, Z_{B_t}, X'_{B_t})$  is some loss function for the source and target domain data in the current mini-batch and  $\lambda_1$  is the trade-off coefficient. For example, if we combine ERM with gradient penalty then  $\hat{L}(W, Z_{B_t}, X'_{B_t}) = \frac{1}{|B_t|} \sum_{k \in B_t} \ell(f_W(X_k), Y_k)$  and  $\ell$  could be the cross-entropy loss. Moreover, if we combine KL guided marginal alignment algorithm [158] with gradient penalty then the objective function is

$$\min_{W, \theta} \frac{1}{|B_t|} \sum_{k \in B_t} \ell(f_W(T_k), Y_k) + \beta_1 \text{D}_{\text{KL}}(P_{T'} || P_T) + \beta_2 \text{D}_{\text{KL}}(P_T || P_{T'}) + \lambda_1 \|g(W, Z_{B_t}, X'_{B_t})\|^2,$$

where  $\theta$  is the parameters of the representation network and the gradient is

$$g(W, Z_{B_t}, X'_{B_t}) = \frac{1}{|B_t|} \sum_{k \in B_t} \nabla_{W, \theta} \ell(f_W(T_k), Y_k) + \beta_1 \nabla_{\theta} \text{D}_{\text{KL}}(P_{T'} || P_T) + \beta_2 \nabla_{\theta} \text{D}_{\text{KL}}(P_T || P_{T'}).$$

In [158], the representation distribution is modelled as an Gaussian distribution, i.e.,  $T \sim \mathcal{N}(\mu_{\theta}, \sigma_{\theta}^2 \mathbf{I}_d | X)$  and  $T' \sim \mathcal{N}(\mu_{\theta}, \sigma_{\theta}^2 \mathbf{I}_d | X')$ . Additionally, let the batch size be  $b = |B_t|$ , the empirical KL divergence is estimated by the mini-batch data, as given in [158],

$$\begin{aligned} & \beta_1 \text{D}_{\text{KL}}(P_{T'} || P_T) + \beta_2 \text{D}_{\text{KL}}(P_T || P_{T'}) \\ & \approx \beta_1 \frac{1}{b} \sum_{k \in B_t} [\log P_{T'_k} - \log P_{T_k}] + \beta_2 \frac{1}{b} \sum_{k \in B_t} [\log P_{T_k} - \log P_{T'_k}] \\ & \approx \beta_1 \frac{1}{b} \sum_{k \in B_t} \left[ \log \frac{1}{b} \sum_{k \in B_t} P_{T'_k | X'_k} - \log \frac{1}{b} \sum_{k \in B_t} P_{T_k | X_k} \right] + \beta_2 \frac{1}{b} \sum_{k \in B_t} \left[ \log \frac{1}{b} \sum_{k \in B_t} P_{T_k | X_k} - \log \frac{1}{b} \sum_{k \in B_t} P_{T'_k | X'_k} \right], \end{aligned}$$

<sup>1</sup>Available at: <https://github.com/facebookresearch/DomainBed>.

<sup>2</sup>Available at: <https://github.com/atuannguyen/kl>.



where  $P_{T_k|X_k} = \mathcal{N}(\mu_\theta, \sigma_\theta^2 \mathbf{I}_d | X_k)$  and  $P_{T'_k|X'_k} = \mathcal{N}(\mu_\theta, \sigma_\theta^2 \mathbf{I}_d | X'_k)$ . To be more precise,  $\mu_\theta$  and  $\sigma_\theta$  are the outputs of the representation network. Since the forward pass requires the sampling of  $T_k$  and  $T'_k$ , we need to use the reparameterization trick [111] for the backward pass.

---

**Algorithm 3** Controlling Label Information

---

**Require:** Source domain labelled dataset  $S$ , Target domain unlabelled dataset  $S'_{X'}$ , Batch size  $b$ , Classification loss function  $\ell_c$ , Marginal alignment loss function  $\ell_r$ , Initial classifier parameter  $\mathbf{w}_0 = \tilde{\mathbf{w}}_0$ , Initial representation network parameter  $\boldsymbol{\theta}_0$ , Learning rate  $\eta$ , Lagrange multiplier  $\lambda_2$

**while**  $\mathbf{w}_t, \theta_t$  not converged **do**

2:   Update iteration:  $t \leftarrow t + 1$

      Sample  $\mathcal{Z}_B = \{\mathbf{z}_i\}_{i=1}^b$  from source domain training set  $S$

4:   Sample  $\mathcal{X}'_B = \{\mathbf{x}'_i\}_{i=1}^b$  from target domain training set  $S'_{X'}$

      Compute distance from the auxiliary classifier  $\mathbf{dis} \leftarrow \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|^2$

6:   Compute marginal alignment loss  $L_r \leftarrow \frac{1}{b} \sum_{i=1}^b \ell_r(\theta_t, \mathbf{z}_i, \mathbf{x}'_i)$

      Compute classification loss  $L_c \leftarrow \frac{1}{b} \sum_{i=1}^b \ell_c(\mathbf{w}_t, \theta_t, \mathbf{z}_i, \mathbf{x}'_i)$

8:   Compute gradient:

$\mathbf{g}_B \leftarrow \nabla(L_c + L_r + \lambda_2 \mathbf{dis})$

      Update parameter:  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \cdot \mathbf{g}_B$ ,  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \cdot \mathbf{g}_B$

10:   Obtain the pseudo labels  $\mathcal{Y}'_B \leftarrow f_{\mathbf{w}_t}(g_{\boldsymbol{\theta}_t}(\mathcal{X}'_B))$

      Compute auxiliary classification loss  $L_a \leftarrow \frac{1}{b} \sum_{i=1}^b \ell_c(\tilde{\mathbf{w}}_t, \theta_t, \mathbf{x}'_i, \mathbf{y}'_i)$

12:   Compute auxiliary classifier gradient:

$\tilde{\mathbf{g}}_B \leftarrow \nabla L_a$

      Update auxiliary classifier parameter:  $\tilde{\mathbf{w}}_{t+1} \leftarrow \tilde{\mathbf{w}}_t - \eta \cdot \tilde{\mathbf{g}}_B$

14: **end while**

---

When we train the model with controlling label information, the objective function becomes

$$\min_W \hat{L}(W, Z_{B_t}, X'_{B_t}) + \lambda_2 \|W - \tilde{W}\|^2,$$

where  $\tilde{W}$  is the auxiliary classifier and  $\lambda_2$  is the trade-off hyperparameter.

Similarly, when we combine KL guided marginal alignment algorithm with controlling label information, then the objective function in every iteration is

$$\min_{W, \theta} \frac{1}{|B_t|} \sum_{k \in B_t} \ell(f_W(T_k), Y_k) + \beta_1 \text{D}_{\text{KL}}(P_{T'} || P_T) + \beta_2 \text{D}_{\text{KL}}(P_T || P_{T'}) + \lambda_2 \|W - \tilde{W}\|^2.$$

In addition, the training objective for the auxiliary classifier is

$$\min_{\widetilde{W}} \frac{1}{|B_t|} \sum_{k \in B_t} \ell(f_{\widetilde{W}}(T'_k), f_W(T'_k)) + \frac{1}{|B_t|} \sum_{k \in B_t} \ell(f_{\widetilde{W}}(T_k), f_W(T_k)). \quad (\text{D.10})$$

In practice, removing the second term would not affect the performance. Note that we need to disable the automatic differentiation of  $T$ ,  $T'$  and  $W$  when executing the backward pass for the auxiliary classifier. The detailed algorithm of controlling label information is given in the next section.

#### D.4.2 Algorithm of Controlling Label Information and Additional Results of ERM-CL

If we only provide the pseudo labels for the target domain data to the auxiliary classifier, i.e. removing the second term in Eq (D.10), the Algorithm 3 is the algorithm for combining any marginal alignment algorithm with controlling label information.

Even without incorporating with the marginal alignment algorithm, e.g., ERM, in which case  $L_r$  is removed, Algorithm 3 still boosts the performance in practice.

Table D.1 shows that **ERM-CL** can overall outperform the basic **ERM** and is close to the performance of **ERM-GP**.

#### D.4.3 Architectures and Hyperparameters

The network architecture in this work is the same as in [70] and [158], where a simple CNN is used.

Other settings are also the same as [70] and [158], for example, each algorithm is trained for 100 epochs. To select the hyperparameters ( $\lambda_1$  and  $\lambda_2$ ) for **ERM-GP**, **ERM-KL**, **KL-GP** and **KL-CL**, we perform random search. Specifically,  $\lambda_1$  is searched between  $[0.1, 0.9]$  and  $\lambda_2$  is searched between  $[10^{-6}, 0.8]$ . Other hyperparameters searching range could be found in the source code of [158].

#### D.4.4 Ablation Study on the Effect of Gradient Penalty Hyperparameter

Our study includes an ablation analysis to investigate the impact of the hyperparameter  $\lambda_1$  in the context of **KL-GP**. Specifically, we conduct experiments on both RotatedM-

Table D.1: RotatedMNIST and Digits Experiments of **ERM-CL**. Results of ERM are reported from [158].

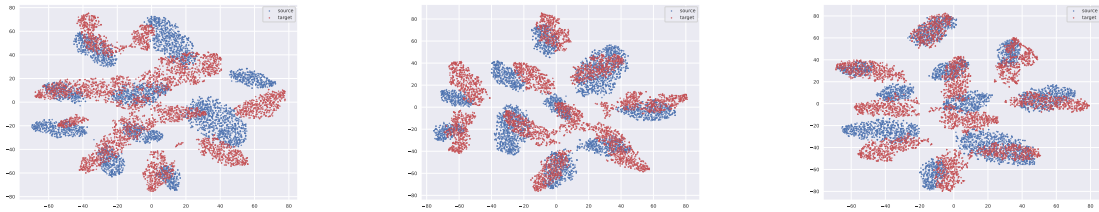
| Method | RotatedMNIST ( $0^\circ$ as source domain) |                                |                                |                                |                                |             | Digits                         |                                |                                |             |
|--------|--------------------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|-------------|--------------------------------|--------------------------------|--------------------------------|-------------|
|        | $15^\circ$                                 | $30^\circ$                     | $45^\circ$                     | $60^\circ$                     | $75^\circ$                     | Ave         | M $\rightarrow$ U              | U $\rightarrow$ M              | S $\rightarrow$ M              | Ave         |
| ERM    | 97.5 $\pm$ 0.2                             | 84.1 $\pm$ 0.8                 | 53.9 $\pm$ 0.7                 | 34.2 $\pm$ 0.4                 | 22.3 $\pm$ 0.5                 | 58.4        | 73.1 $\pm$ 4.2                 | 54.8 $\pm$ 6.2                 | 65.9 $\pm$ 1.4                 | 64.6        |
| ERM-GP | <b>97.5<math>\pm</math>0.1</b>             | <b>86.2<math>\pm</math>0.5</b> | <b>62.0<math>\pm</math>1.9</b> | <b>34.8<math>\pm</math>2.1</b> | <b>26.1<math>\pm</math>1.2</b> | <b>61.2</b> | <b>91.3<math>\pm</math>1.6</b> | <b>72.7<math>\pm</math>4.2</b> | 68.4 $\pm$ 0.2                 | 77.5        |
| ERM-CL | 97.3 $\pm$ 0.1                             | 84.1 $\pm$ 0.1                 | 56.9 $\pm$ 2.5                 | 34.2 $\pm$ 1.9                 | 25.5 $\pm$ 1.6                 | 59.6        | 88.9 $\pm$ 0.4                 | 71.2 $\pm$ 3.6                 | <b>73.5<math>\pm</math>1.4</b> | <b>77.9</b> |

Table D.2: Ablation study on Effect of Gradient Penalty Hyperparameter.

| $\lambda_1$                    | 0              | 0.1            | 0.3            | 0.5            |
|--------------------------------|----------------|----------------|----------------|----------------|
| $0^\circ \rightarrow 60^\circ$ | 75.5 $\pm$ 2.4 | 88.0 $\pm$ 8.1 | 82.8 $\pm$ 5.8 | 80.1 $\pm$ 3.7 |
| S $\rightarrow$ M              | 92.5 $\pm$ 0.9 | 93.6 $\pm$ 1.2 | 93.8 $\pm$ 1.1 | 93.1 $\pm$ 1.7 |

NIST and Digits datasets, where the source and target domains are set to  $0^\circ/60^\circ$  and SVHN/MNIST, respectively. Table D.2 summarizes the results. It is worth noting that setting  $\lambda_1$  to zero effectively reduces **KL-GP** to **KL**, and our results confirm the efficacy of including the gradient penalty term in **KL-GP**.

#### D.4.5 Visualization Results



(a) Visualization of KL (b) Visualization of KL-GP (c) Visualization of KL-CL

Figure D.2: Visualization results of representations obtained by using t-SNE. The source domain (blue points) is SVHN and the target domain (red points) is MNIST.

To visualize the representations of models trained using **KL**, **KL-GP**, and **KL-CL**, we employ t-SNE [205]. Figure D.2 displays the visualization results when SVHN is used as

Table D.3: VisDA17 experiments. Results of baselines are reported directly from [158].

| Method | Synthetic $\rightarrow$ Real   |
|--------|--------------------------------|
| ERM    | 39.1 $\pm$ 0.5                 |
| DANN   | 57.7 $\pm$ 1.3                 |
| MMD    | 62.8 $\pm$ 1.1                 |
| CORAL  | 39.5 $\pm$ 4.5                 |
| WD     | 38.9 $\pm$ 4.8                 |
| KL     | 70.6 $\pm$ 0.5                 |
| KL-GP  | <b>71.9<math>\pm</math>0.7</b> |
| KL-CL  | 71.3 $\pm$ 0.4                 |

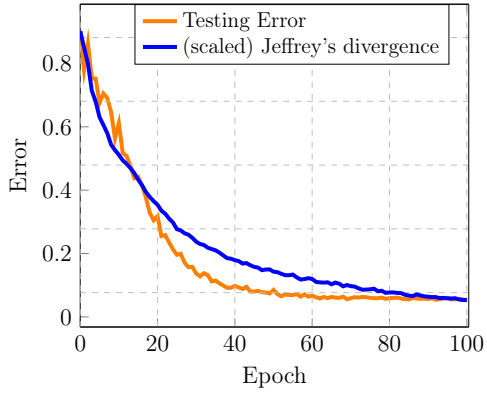
the source domain and MNIST as the target domain. Our findings indicate that the incorporation of additional regularizers yields a slight improvement in representation alignment. However, it is essential to note that these regularization terms are primarily designed to enhance the performance of the classifier network, rather than the representation network.

#### D.4.6 Results on VisDA17

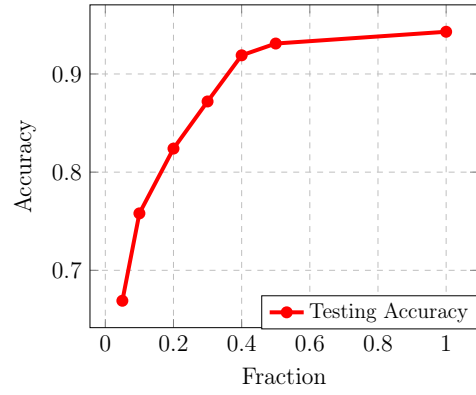
We also conduct experiments on the VisDA17 dataset [167], which is a real-world classification task with 280K images from 12 classes. Particularly, the source domain contains synthetic images and the target domain contains real images. Table D.3 presents our experimental results. Notably, our regularization techniques, namely **KL-GP** and **KL-CL** are still capable of improving the performance of the KL guided marginal alignment algorithm to some extent.

#### D.4.7 Dynamics of Jeffrey’s divergence

The representation space version of Corollary 6.4.2 suggests that a small Jeffrey’s divergence can lead to a low testing error. Figure D.3a demonstrates that the dynamic of Jeffrey’s divergence, as computed in the representation space, can effectively characterize the evolution of the testing error throughout the training phase. Additionally, Figure D.3b reveals that the number of target data used has an impact on testing performance. Specifically, when less than half of the available unlabelled target data is used, performance



(a) Training Dynamic



(b) Testing Performance vs. # of Target Data

Figure D.3: **KL** on **S**→**M**. The left figure is the comparison of the Jeffrey's divergence in the representation space and the testing error. The right figure is the evolution of testing accuracy with respect to the different fraction of unlabelled target data used for training.

increases with the number of data. However, when more than half of the unlabelled target data is used, there is only marginal improvement on performance.

# Appendix E

## Appendix of Chapter 7

### E.1 Useful Facts

In Chapter 7, there are some equivalent forms of the generalization error we will study, e.g., Eq. (7.2) and Eq. (7.5) in the main text, which are stated in the following lemma.

**Lemma E.1.1.** *Let  $W_i = \widetilde{W}_{i,U_i}$  and  $\overline{W}_i = \widetilde{W}_{i,\overline{U}_i}$ . For any learning algorithm  $\mathcal{A}$ , the following equations hold*

$$\mathcal{E}_\mu(\mathcal{A}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\widetilde{Z}_i^+, \widetilde{W}_i} \left[ \ell(\widetilde{W}_i^-, \widetilde{Z}_i^+) - \ell(\widetilde{W}_i^+, \widetilde{Z}_i^+) \right], \quad (\text{E.1})$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\widetilde{W}_i} \left[ \mathbb{E}_{\widehat{Z}_i, U_i | \widetilde{W}_i} \left[ (-1)^{U_i} \left( \ell(\widetilde{W}_i^-, \widehat{Z}_i) - \ell(\widetilde{W}_i^+, \widehat{Z}_i) \right) \right] \right], \quad (\text{E.2})$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\widetilde{Z}_i^+} \left[ \mathbb{E}_{W_i, \overline{W}_i, U_i | \widetilde{Z}_i^+} \left[ (-1)^{U_i} \left( \ell(\overline{W}_i, \widetilde{Z}_i^+) - \ell(W_i, \widetilde{Z}_i^+) \right) \right] \right]. \quad (\text{E.3})$$

*Proof.* This lemma is a consequence of Lemma 7.3.1, with further utilizing some symmetric

properties. Recall Eq. (7.1) in Lemma 7.3.1,

$$\begin{aligned}
\mathcal{E}_\mu(\mathcal{A}) &= \mathbb{E}_{\tilde{Z}_{[n]}^+, \tilde{Z}_{[n]}^-} \left[ \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{\tilde{W}_i^- | \tilde{Z}_{[n]}^+, \tilde{Z}_i^-} \ell(\tilde{W}_i^-, \tilde{Z}_i^+) - \mathbb{E}_{\tilde{W}_i^+ | \tilde{Z}_{[n]}^-} \ell(\tilde{W}_i^+, \tilde{Z}_i^-) \right] \right], \\
&= \mathbb{E}_{\tilde{Z}_{[n]}^+, \tilde{W}} \left[ \frac{1}{n} \sum_{i=1}^n \left[ \ell(\tilde{W}_i^-, \tilde{Z}_i^+) - \ell(\tilde{W}_i^+, \tilde{Z}_i^-) \right] \right], \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i^+, \tilde{W}_i} \left[ \ell(\tilde{W}_i^-, \tilde{Z}_i^+) - \ell(\tilde{W}_i^+, \tilde{Z}_i^-) \right].
\end{aligned}$$

Note that Eq. (7.2) in the main text is from the second equation above, which is used to derive individual IOMI bounds in Section 7.4.

Similar to the standard setting for CMI bounds, where the role of each  $\tilde{Z}_i^+$  and  $\tilde{Z}_i^-$  can be exchanged, a key observation here is that for each  $i$ ,  $\tilde{W}_i^+$  and  $\tilde{W}_i^-$  can also be exchanged arbitrarily. That is to say,

$$\mathcal{E}_\mu(\mathcal{A}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i^-, \tilde{W}_i} \left[ \ell(\tilde{W}_i^+, \tilde{Z}_i^-) - \ell(\tilde{W}_i^-, \tilde{Z}_i^+) \right] \quad (\text{E.4})$$

also holds true. Notice that we do not change the definitions of any the random variable, e.g.,  $\tilde{W}^+ = \mathcal{A}(\tilde{Z}_{[n]}^+, R)$  and  $\tilde{W}_i^- = \mathcal{A}(\tilde{Z}_{[n] \sim i}^+, R)$ .

What differs from the standard CMI is that the roles of the whole sequences  $\tilde{Z}_{[n]}^+$  and  $\tilde{Z}_{[n]}^-$  are not exchangeable with each other. Here, when we exchange each  $\tilde{Z}_i^+$  and  $\tilde{Z}_i^-$ , we need to keep the other positions in  $S$  unchanged.

By introducing  $U_i \sim \text{Unif}(\{0, 1\})$ , we have

$$\begin{aligned}
\mathcal{E}_\mu(\mathcal{A}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i^+, \tilde{W}_i} \left[ \ell(\tilde{W}_i^-, \tilde{Z}_i^+) - \ell(\tilde{W}_i^+, \tilde{Z}_i^-) \right], \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i, \tilde{W}_i, U_i} \left[ \ell(\tilde{W}_{i, \bar{U}_i}, \tilde{Z}_{i, U_i}) - \ell(\tilde{W}_{i, U_i}, \tilde{Z}_{i, \bar{U}_i}) \right].
\end{aligned}$$

To obtain Eq. (E.2), notice that  $\hat{Z}_i = \tilde{Z}_{i, U_i}$ , we have

$$\begin{aligned}
\mathcal{E}_\mu(\mathcal{A}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{Z}_i, \tilde{W}_i, U_i} \left[ \ell(\tilde{W}_{i, \bar{U}_i}, \hat{Z}_i) - \ell(\tilde{W}_{i, U_i}, \hat{Z}_i) \right] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{Z}_i, \tilde{W}_i, U_i} \left[ (-1)^{U_i} \left( \ell(\tilde{W}_i^-, \hat{Z}_i) - \ell(\tilde{W}_i^+, \hat{Z}_i) \right) \right]. \quad (\text{E.5})
\end{aligned}$$

This, as we have already seen in Eq. (7.5) in the main text, is used to derive hypotheses-conditioned CMI bounds in Section 7.5. It's easy to see that when  $U_i = 0$ , Eq. (E.5) becomes Eq. (E.1), and when  $U_i = 1$ , we obtain Eq. (E.4) via Eq. (E.5).

To obtain Eq. (E.3), we let  $W_i = \widetilde{W}_{i,U_i}$ ,  $\overline{W}_i = \widetilde{W}_{i,\overline{U}_i}$ , and fix  $\hat{Z}_i = \widetilde{Z}_i^+$ . Similarly,

$$\mathcal{E}_\mu(\mathcal{A}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\widetilde{Z}_i^+} \left[ \mathbb{E}_{W_i, \overline{W}_i, U_i | \widetilde{Z}_i^+} \left[ (-1)^{U_i} \left( \ell(\overline{W}_i, \widetilde{Z}_i^+) - \ell(W_i, \widetilde{Z}_i^+) \right) \right] \right].$$

This is used to derive supsample-conditioned CMI bounds in Section 7.5. It's easy to see that both  $U_i = 0$  and  $U_i = 1$  will give us Eq. (E.1).  $\square$

## E.2 Further Elaborations on SCH Stability

We note that the reason we introduce four types of SCH stability in Definition 7.3.1 is that solely using  $\beta_2$  in our bounds might be too loose, as it considers the supremum over all sources of randomness. By incorporating SCH stabilities, we aim to demonstrate that theoretically, we can achieve significantly tighter stability parameters.

The basic set up is as follows. Assume a random sample  $S$  gives rise to  $W$ . For each  $Z_i \in S$ , we construct  $S^i$  by replacing  $Z_i$  with another independently drawn instance; call training result  $W^i$ , the neighbor of  $W$ .

In a),  $\gamma_1$ -SCH-A stability measures the difference between the loss of  $w$  and the expected loss of its neighbor  $W^i$  at a worst  $z$  and the worst possible  $w$ . While in (b),  $\gamma_2$ -SCH-B stability measures the square of this difference, not in the worst case, but in an average case, where the average is over an independently  $Z'$  for the loss evaluation, the training sample, and the algorithm randomness. Since “average is smaller than worst”,  $\gamma_2 \leq \gamma_1$ .

In c), we consider the difference between the loss of  $W$  and the loss of its neighbor when evaluated at the worst possible  $Z_i$  that when included in  $S$  gives rise to  $W$ . The expected value of this difference is  $\gamma_3$ -SCH-C stability.

In d),  $\gamma_4$ -SCH-D stability measures the expected squared difference between the loss of  $W$  and the loss of its neighbor when evaluated at  $Z_i$  (a member of  $S$ ). For a similar “average smaller worst” reason, one expects that  $\gamma_4 \leq \gamma_3$ .

We expect that  $\gamma_2, \gamma_3$ , and  $\gamma_4$  are all smaller than  $\beta_1$ . This is because in  $\beta_1$ , we consider the worst evaluated instance, whereas in the other cases, we take the expectation over all instances. Additionally, in Theorem 7.5.1, we expect that  $\mathbb{E}_{\widetilde{W}_i} \Delta_1(\widetilde{W}_i)^2 \leq \beta_1^2$ , this is



because  $\beta_1$ -stability holds for all the possible  $s$  and  $s^i$ , namely it holds for all the  $(w, w^i)$  pair (that shares the same randomness) while in  $\mathbb{E}_{\widetilde{W}_i} \Delta_1(\widetilde{W}_i)^2$ , we take the expectation of these pairs.

We expect  $\gamma_2 \leq \gamma_4$  due to the following reason: first by Jensen's inequality, we have  $\mathbb{E}_{S,R,Z'} [\ell(W, Z') - \mathbb{E}_{W^i|W} \ell(W^i, Z')]^2 \leq \mathbb{E}_{W,W^i,Z'} [\ell(W, Z') - \ell(W^i, Z')]^2$ , then since  $Z'$  is an independent of both  $W$  and  $W'$ ,  $Z'$  can be regarded as a testing point for both  $W$  and  $W'$ , we could expect that the expectation of  $\ell(W, Z') - \ell(W^i, Z')$  is small. While in  $\mathbb{E}_{S,Z'_i,R} [\ell(W, Z_i) - \ell(W^i, Z_i)]^2$ ,  $Z_i$  is a training point for obtaining  $W$ , so  $\ell(W, Z_i)$  could be small in general, and  $Z_i$  is a testing point for  $W^i$ . Therefore, it is reasonable to expect  $\mathbb{E}_{W,W^i,Z'} [\ell(W, Z') - \ell(W^i, Z')]^2 \leq \mathbb{E}_{S,Z'_i,R} [\ell(W, Z_i) - \ell(W^i, Z_i)]^2$ , namely  $\gamma_2 \leq \gamma_4$ .

As a concrete example, let  $\ell$  be zero-one loss and assume  $\mathcal{A}$  is an interpolating algorithm and randomly makes predictions for unseen data. By Jensen's inequality,  $\gamma_2^2 \leq \mathbb{E}_{W,W^i,Z'} [\ell(W, Z') - \ell(W^i, Z')]^2 = \mathbb{E}_{W,Z'} [\ell(W, Z')] - 2\mathbb{E}_{W,W^i,Z'} [(\ell(W, Z')\ell(W^i, Z'))] + \mathbb{E}_{W^i,Z'} [\ell(W^i, Z')]^2$ , where we use  $\ell^2 = \ell$  for zero-one loss. Since  $Z'$  is an unseen data for both  $W$  and  $W^i$ , we have  $\gamma_2^2 \leq \mathbb{E}_{W^i,Z'} [\ell(W^i, Z')]^2 + \frac{1}{2} - \frac{1}{2} = \mathbb{E}_{W^i,Z'} [\ell(W^i, Z')]^2$ . While in this case  $\gamma_4^2 = \mathbb{E}_{W^i,Z_i} [\ell(W^i, Z_i)]^2$  so  $\gamma_2 \leq \gamma_4$ .

## E.3 Omitted Proofs and Additional Discussions in Section 7.4

### E.3.1 Proof of Theorem 7.4.1

*Proof.* Let  $g(\tilde{w}^+, \tilde{z}_i^+) = \mathbb{E}_{\widetilde{W}_i^-|\tilde{w}^+} [\ell(\widetilde{W}_i^-, \tilde{z}_i^+)] - \ell(\tilde{w}^+, \tilde{z}_i^+)$  be the average loss difference between  $\tilde{w}^+$  and its neighboring hypothesis, and let  $f = t \cdot g$  for  $t > 0$  in Lemma 2.4.1. Let  $\tilde{Z}_i^{+'}$  be an independent copy of  $\tilde{Z}_i^+$ , then

$$\mathbb{E}_{\widetilde{W}^+, \tilde{Z}_i^+} [g(\widetilde{W}^+, \tilde{Z}_i^+)] \leq \inf_{t>0} \frac{I(\widetilde{W}^+; \tilde{Z}_i^+) + \log \mathbb{E}_{\widetilde{W}^+, \tilde{Z}_i^{+'}} [e^{tg(\widetilde{W}^+, \tilde{Z}_i^{+'})}]}{t}. \quad (\text{E.6})$$

Since  $\tilde{Z}_i^{+'}$  is independent of both  $\widetilde{W}_i^-$  and  $\widetilde{W}^+$ , and  $\widetilde{W}_i^-$  and  $\widetilde{W}^+$  are identically distributed, we have

$$\mathbb{E}_{\widetilde{W}^+, \tilde{Z}_i^{+'}} [g(\widetilde{W}^+, \tilde{Z}_i^{+'})] = \mathbb{E}_{\widetilde{W}_i^-, \tilde{Z}_i^{+'}} [\ell(\widetilde{W}_i^-, \tilde{Z}_i^{+'})] - \mathbb{E}_{\widetilde{W}^+, \tilde{Z}_i^{+'}} [\ell(\widetilde{W}^+, \tilde{Z}_i^{+'})] = 0.$$

By the definition of  $\gamma_1$ -SCH-A stability,

$$\sup_{\tilde{w}^+, z} \left| \mathbb{E}_{\tilde{W}_i^- | \tilde{w}^+} \left[ \ell(\tilde{W}_i^-, z) \right] - \ell(\tilde{w}^+, z) \right| \leq \gamma_1,$$

so  $g(\tilde{W}^+, \tilde{Z}_i^{+'})$  is a zero-mean random variable bounded in  $[-\gamma_1, \gamma_1]$ . By Lemma 2.4.2, we have

$$\begin{aligned} \mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^+} \left[ g(\tilde{W}^+, \tilde{Z}_i^+) \right] &\leq \inf_{t>0} \frac{I(\tilde{W}^+; \tilde{Z}_i^+) + \log \mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^{+'}} \left[ e^{tg(\tilde{W}^+, \tilde{Z}_i^{+'})} \right]}{t} \\ &\leq \inf_{t>0} \frac{I(\tilde{W}^+; \tilde{Z}_i^+) + \frac{t^2 \gamma_1^2}{2}}{t} \\ &= \sqrt{2\gamma_1^2 I(\tilde{W}^+; \tilde{Z}_i^+)}, \end{aligned}$$

where the last equality is obtained by optimizing the bound over  $t$ , i.e. let  $t = \sqrt{\frac{I(\tilde{W}^+; \tilde{Z}_i^+)}{2\gamma_1^2}}$ .

Recall Eq. (E.1) in Lemma E.1.1 and applying Jensen's inequality to the absolute function, the first bound is then obtained by

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^+} \left[ g(\tilde{W}^+, \tilde{Z}_i^+) \right] \right| \leq \frac{\gamma_1}{n} \sum_{i=1}^n \sqrt{2I(\tilde{W}^+; \tilde{Z}_i^+)},$$

Furthermore, by the chain rule of mutual information,

$$I(\tilde{W}_i^-; \tilde{Z}_i^+ | \tilde{W}^+) + I(\tilde{W}^+; \tilde{Z}_i^+) = I(\tilde{W}^+; \tilde{Z}_i^+ | \tilde{W}_i^-) + I(\tilde{W}_i^-; \tilde{Z}_i^+). \quad (\text{E.7})$$

Notice that  $I(\tilde{W}_i^-; \tilde{Z}_i^+) = 0$  in the RHS, we have

$$I(\tilde{W}^+; \tilde{Z}_i^+) \leq I(\tilde{W}^+; \tilde{Z}_i^+ | \tilde{W}_i^-),$$

which will give us the second bound. This concludes the proof.  $\square$

**Remark E.3.1** (Comparison with Mutual Information Stability [173, 82]). *To compare with the mutual information stability  $I(\tilde{W}^+; \tilde{Z}_i^+ | \tilde{Z}_{[n] \setminus i}^+)$ , recall Eq. (E.7):  $I(\tilde{W}^+; \tilde{Z}_i^+ | \tilde{W}^-) = I(\tilde{W}_i^-; \tilde{Z}_i^+ | \tilde{W}^+) + I(\tilde{W}^+; \tilde{Z}_i^+)$ , and similarly we also have  $I(\tilde{W}^+; \tilde{Z}_i^+ | \tilde{Z}_{[n] \setminus i}^+) = I(\tilde{Z}_{[n] \setminus i}^+; \tilde{Z}_i^+ | \tilde{W}^+) + I(\tilde{W}^+; \tilde{Z}_i^+)$ .*

Thus, we only need to compare  $I(\widetilde{W}_i^-; \widetilde{Z}_i^+ | \widetilde{W}^+)$  and  $I(\widetilde{Z}_{[n]\setminus i}^+; \widetilde{Z}_i^+ | \widetilde{W}^+)$ . Notice that for a deterministic  $\mathcal{A}$ , we have  $I(\widetilde{W}_i^-; \widetilde{Z}_i^+ | \widetilde{W}^+) \leq I(\widetilde{Z}_{[n]\setminus i}^+; \widetilde{Z}_i^-; \widetilde{Z}_i^+ | \widetilde{W}^+)$ . Since  $\widetilde{Z}_i^- \perp\!\!\!\perp (\widetilde{W}^+, \widetilde{Z}_{[n]}^+)$ , we further have  $I(\widetilde{Z}_{[n]\setminus i}^+; \widetilde{Z}_i^-; \widetilde{Z}_i^+ | \widetilde{W}^+) = I(\widetilde{Z}_{[n]\setminus i}^+; \widetilde{Z}_i^+ | \widetilde{W}^+)$ , which gives us the desired result:

$$I(\widetilde{W}_i^+; \widetilde{Z}_i^+ | \widetilde{W}^-) \leq I(\widetilde{W}^+; \widetilde{Z}_i^+ | \widetilde{Z}_{[n]\setminus i}^+).$$

### E.3.2 Proof of Theorem 7.4.2

*Proof.* The proof is nearly the same to the proof of Theorem 7.4.1, except that now the randomness of the algorithm is given for each DV auxiliary function, so the randomness of  $\widetilde{W}_i$  is completely controlled by  $\widetilde{Z}$ .

Let  $g(\widetilde{w}^+, \widetilde{z}_i^+, r) = \mathbb{E}_{\widetilde{W}_i^- | \widetilde{w}^+, r} [\ell(\widetilde{W}_i^-, \widetilde{z}_i^+)] - \ell(\widetilde{w}^+, \widetilde{z}_i^+)$  and let  $f = t \cdot g$  for  $t > 0$  in Lemma 2.4.1. Let  $\widetilde{Z}_i^{+'}$  be an independent copy of  $\widetilde{Z}_i^+$ , then

$$\mathbb{E}_{\widetilde{W}^+, \widetilde{Z}_i^+ | r} [g(\widetilde{W}^+, \widetilde{Z}_i^+, r)] \leq \inf_{t>0} \frac{I(\widetilde{W}^+; \widetilde{Z}_i^+ | R = r) + \log \mathbb{E}_{\widetilde{W}^+, \widetilde{Z}_i^{+'} | r} [e^{tg(\widetilde{W}^+, \widetilde{Z}_i^{+'}, r)}]}{t}.$$

Notice that

$$\mathbb{E}_{\widetilde{W}^+, \widetilde{Z}_i^{+'} | r} [g(\widetilde{W}^+, \widetilde{Z}_i^{+'}, r)] = \mathbb{E}_{\widetilde{W}_i^-, \widetilde{Z}_i^{+'} | r} [\ell(\widetilde{W}_i^-, \widetilde{Z}_i^{+'})] - \mathbb{E}_{\widetilde{W}_i^+, \widetilde{Z}_i^{+'} | r} [\ell(\widetilde{W}_i^+, \widetilde{Z}_i^{+'})] = 0$$

still holds since  $\widetilde{Z}_i^+$  and  $\widetilde{Z}_i^-$  are i.i.d. drawn.

Thus,  $g(\widetilde{W}^+, \widetilde{Z}_i^{+'}, r)$  is a zero-mean random variable bounded in  $[-\beta_2, \beta_2]$ . By Lemma 2.4.2, the remaining part is routine:

$$\mathbb{E}_{\widetilde{W}^+, \widetilde{Z}_i^+ | r} [g(\widetilde{W}^+, \widetilde{Z}_i^+, r)] \leq \sqrt{2\beta_2^2 I(\widetilde{W}^+; \widetilde{Z}_i^+ | R = r)}.$$

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{\widetilde{W}^+, \widetilde{Z}_i^+, R} [g(\widetilde{W}^+, \widetilde{Z}_i^+, R)] \right| \leq \frac{\beta_2}{n} \sum_{i=1}^n \mathbb{E}_R \sqrt{2I^R(\widetilde{W}^+; \widetilde{Z}_i^+)},$$

This completes the proof. □

### E.3.3 Proof of Theorem 7.4.3

*Proof.* Let  $h(x) = \frac{e^x - x - 1}{x^2}$  be the Bernstein function. Similar to the proof of Theorem 7.4.1, we let  $g(\tilde{w}^+, \tilde{z}_i^+) = \mathbb{E}_{\tilde{W}_i^- | \tilde{w}^+} [\ell(\tilde{W}_i^-, \tilde{z}_i^+)] - \ell(\tilde{w}^+, \tilde{z}_i^+)$ . We have already known that  $\mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^{+'}} [g(\tilde{W}^+, \tilde{Z}_i^{+'})] = 0$  and  $|g(\tilde{W}^+, \tilde{Z}_i^{+'})| \leq \gamma_1$ . By Lemma 2.4.4,

$$\begin{aligned} \log \mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^{+'}} [e^{tg(\tilde{W}^+, \tilde{Z}_i^{+'})}] &\leq h(\gamma_1 t) t^2 \mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^{+'}} \left[ \left( \mathbb{E}_{\tilde{W}_i^- | \tilde{w}^+} [\ell(\tilde{W}_i^-, \tilde{Z}_i^{+'})] - \ell(\tilde{W}^+, \tilde{Z}_i^{+'}) \right)^2 \right] \\ &\leq h(\gamma_1 t) t^2 \gamma_2^2, \end{aligned}$$

where the second inequality is by the definition of  $\gamma_2$ -SCH-B stability.

Plugging the above into Eq. (E.6),

$$\begin{aligned} \mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^+} [g(\tilde{W}^+, \tilde{Z}_i^+)] &\leq \inf_{t>0} \frac{I(\tilde{W}^+; \tilde{Z}_i^+) + \log \mathbb{E}_{\tilde{W}^+, \tilde{Z}_i^{+'}} [e^{tg(\tilde{W}^+, \tilde{Z}_i^{+'})}]}{t} \\ &\leq \inf_{t>0} \frac{I(\tilde{W}^+; \tilde{Z}_i^+)}{t} + h(\gamma_1 t) t \gamma_2^2. \end{aligned}$$

Usually we have  $\gamma_2^2 \leq \gamma_1^2 \leq \gamma_1$ , we let  $t = 1/\gamma_1$ , then

$$h(\gamma_1 t) t \gamma_2^2 = \frac{h(1) \gamma_2^2}{\gamma_1} \approx 0.72 \frac{\gamma_2^2}{\gamma_1}.$$

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\gamma_1}{n} \sum_{i=1}^n I(\tilde{W}^+; \tilde{Z}_i^+) + \frac{0.72 \gamma_2^2}{\gamma_1}.$$

This concludes the proof.  $\square$

## E.4 Omitted Proofs in Section 7.5

### E.4.1 Proof of Theorem 7.5.1

*Proof.* We now prove the first bound. Let  $g(\tilde{w}_i, \hat{z}_i, u_i) = (-1)^{u_i} (\ell(\tilde{w}_i^-, \hat{z}_i) - \ell(\tilde{w}_i^+, \hat{z}_i))$ . By Lemma 2.4.1, we have

$$\mathbb{E}_{\hat{Z}_i, U_i | \tilde{w}_i} [g(\tilde{w}_i, \hat{Z}_i, U_i)] \leq \inf_{t>0} \frac{I(\hat{Z}_i; U_i | \tilde{W}_i = \tilde{w}_i) + \log \mathbb{E}_{\hat{Z}_i, U_i' | \tilde{w}_i} [e^{tg(\tilde{w}_i, \hat{Z}_i, U_i')}] }{t}. \quad (\text{E.8})$$

Since  $U'_i \perp\!\!\!\perp \widehat{Z}_i$ , we have  $\mathbb{E}_{U'_i} [g(\tilde{w}_i, \hat{z}_i, U'_i)] = \mathbb{E}_{U'_i} [(-1)^{U'_i} (\ell(\tilde{w}_i^-, \hat{z}_i) - \ell(\tilde{w}_i^+, \hat{z}_i))] = 0$  for any  $\tilde{w}_i$  and  $\hat{z}_i$ . Ergo,

$$\mathbb{E}_{\widehat{Z}_i|\tilde{w}_i} \left[ \mathbb{E}_{U'_i} \left[ g(\tilde{w}_i, \widehat{Z}_i, U'_i) \right] \right] = 0.$$

By the definition of  $\Delta_1(\tilde{w}_i)$ ,

$$\left| g(\tilde{w}_i, \widehat{Z}_i, U'_i) \right| = \left| \ell(\tilde{w}_i^-, \widehat{Z}_i) - \ell(\tilde{w}_i^+, \widehat{Z}_i) \right| \leq \sup_{z_i \in \mathcal{Z}_{\tilde{w}_i}} |\ell(\tilde{w}_i^-, z_i) - \ell(\tilde{w}_i^+, z_i)| \leq \Delta_1(\tilde{w}_i).$$

Thus,  $g(\tilde{w}_i, \widehat{Z}_i, U'_i)$  is a zero-mean random variable bounded in  $[-\Delta_1(\tilde{w}_i), \Delta_1(\tilde{w}_i)]$  for a fixed  $\tilde{w}_i$ . By Lemma 2.4.2, we have

$$\mathbb{E}_{\widehat{Z}_i, U'_i|\tilde{w}_i} \left[ e^{tg(\tilde{w}_i, \widehat{Z}_i, U'_i)} \right] \leq e^{\frac{t^2 \Delta_1(\tilde{w}_i)^2}{2}}.$$

Plugging the above into Eq. (E.8),

$$\begin{aligned} \mathbb{E}_{\widehat{Z}_i, U_i|\tilde{w}_i} \left[ g(\tilde{w}_i, \widehat{Z}_i, U_i) \right] &\leq \inf_{t>0} \frac{I(\widehat{Z}_i; U_i | \widetilde{W}_i = \tilde{w}_i) + \frac{t^2 \Delta_1(\tilde{w}_i)^2}{2}}{t} \\ &= \Delta_1(\tilde{w}_i) \sqrt{2I(\widehat{Z}_i; U_i | \widetilde{W}_i = \tilde{w}_i)}. \end{aligned} \quad (\text{E.9})$$

Recall Eq. (E.3) in Lemma E.1.1 and by Jensen's inequality for the absolute function, the first bound is obtained:

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\widetilde{W}_i} \left[ \Delta_1(\widetilde{W}_i) \sqrt{2I(\widehat{Z}_i; U_i | \widetilde{W}_i)} \right]. \quad (\text{E.10})$$

To prove the second bound, we return to Eq. (E.9), and take expectation over  $\widetilde{W}_i$  first. By Jensen's inequality,

$$\begin{aligned} \mathbb{E}_{\widehat{Z}_i, U_i, \widetilde{W}_i} \left[ g(\widetilde{W}_i, \widehat{Z}_i, U_i) \right] &\leq \inf_{t>0} \frac{I(\widehat{Z}_i; U_i | \widetilde{W}_i) + \frac{t^2 \mathbb{E}_{\widetilde{W}_i} [\Delta(\widetilde{W}_i)^2]}{2}}{t} \\ &= \sqrt{2\mathbb{E}_{\widetilde{W}_i} [\Delta(\widetilde{W}_i)^2]} I(\widehat{Z}_i; U_i | \widetilde{W}_i). \end{aligned} \quad (\text{E.11})$$

Therefore, we have the second bound as below

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\mathbb{E}_{\widetilde{W}_i} [\Delta(\widetilde{W}_i)^2]} I(\widehat{Z}_i; U_i | \widetilde{W}_i). \quad (\text{E.12})$$

For the second part of Theorem 7.5.1, notice that it's valid to let  $\gamma_3 = \mathbb{E}_{\widetilde{W}_i} [\Delta(\widetilde{W}_i)]$ , then recall Eq. (E.10),

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\widetilde{W}_i} \left[ \Delta(\widetilde{W}_i) \sqrt{2I(\widetilde{W}_i; \widehat{Z}_i; U_i)} \right] \leq \frac{\sqrt{2}\gamma_3}{n} \sum_{i=1}^n \sqrt{\sup_{\widetilde{w}_i \in (\mathcal{W}_s)^2_{s \in \mathcal{Z}^n}} I(\widetilde{w}_i; \widehat{Z}_i; U_i)}.$$

This completes the proof.  $\square$

## E.4.2 Proof of Theorem 7.5.2

*Proof.* The proof is similar to [77, Theorem 2.1]. By the chain rule,

$$I(\widehat{Z}_i; U_i, \widetilde{W}_i) = I(\widehat{Z}_i; U_i | \widetilde{W}_i) + I(\widehat{Z}_i; \widetilde{W}_i). \quad (\text{E.13})$$

Since  $H(\widehat{Z}_i | U_i, \widetilde{W}_i) = H(\widehat{Z}_i | W_i, U_i, \widetilde{W}_i) = H(\widehat{Z}_i | W_i)$ , we have  $I(\widehat{Z}_i; U_i, \widetilde{W}_i) = H(\widehat{Z}_i) - H(\widehat{Z}_i | U_i, \widetilde{W}_i) = H(\widehat{Z}_i) - H(\widehat{Z}_i | W_i) = I(\widehat{Z}_i; W_i)$ . Thus,  $I(\widehat{Z}_i; U_i, \widetilde{W}_i) = I(\widehat{Z}_i; W_i)$ . Recall Eq. (E.13) and by the non-negativity of mutual information, we have  $I(\widehat{Z}_i; U_i | \widetilde{W}_i) \leq I(W_i; \widehat{Z}_i)$ . Note that  $I(W_i; \widehat{Z}_i) = I(\widetilde{W}_i^+; \widetilde{Z}_i^+) = I(W; Z_i)$ . This completes the proof.  $\square$

## E.4.3 Proof of Theorem 7.5.3

*Proof.* We first return to Eq. (E.8) in the previous proof, and we have already known that  $g(\widetilde{w}_i, \widehat{Z}_i, U'_i)$  is a zero-mean random variable bounded in  $[-\Delta_1(\widetilde{w}_i), \Delta_1(\widetilde{w}_i)]$  for a fixed  $\widetilde{w}_i$ .

By Lemma 2.4.4, we have

$$\begin{aligned} \log \mathbb{E}_{\widehat{Z}_i, U'_i | \widetilde{w}_i} \left[ e^{tg(\widetilde{w}_i, \widehat{Z}_i, U'_i)} \right] &\leq h(\Delta_1(\widetilde{w}_i)t) t^2 \mathbb{E}_{\widehat{Z}_i, U'_i | \widetilde{w}_i} \left[ g(\widetilde{w}_i, \widehat{Z}_i, U'_i)^2 \right] \\ &= h(\Delta_1(\widetilde{w}_i)t) t^2 \mathbb{E}_{\widehat{Z}_i | \widetilde{w}_i} \left[ \left( \ell(\widetilde{w}_i^-, \widehat{Z}_i) - \ell(\widetilde{w}_i^+, \widehat{Z}_i) \right)^2 \right]. \end{aligned}$$

Plugging the above into Eq. (E.8),

$$\mathbb{E}_{\widehat{Z}_i, U_i | \widetilde{w}_i} \left[ g(\widetilde{w}_i, \widehat{Z}_i, U_i) \right] \leq \inf_{t>0} \frac{I(\widehat{Z}_i; U_i | \widetilde{W}_i = \widetilde{w}_i)}{t} + h(\Delta_1(\widetilde{w}_i)t) t \mathbb{E}_{\widehat{Z}_i | \widetilde{w}_i} \left[ \left( \ell(\widetilde{w}_i^-, \widehat{Z}_i) - \ell(\widetilde{w}_i^+, \widehat{Z}_i) \right)^2 \right]. \quad (\text{E.14})$$

Let  $t = \frac{1}{\Delta_1(\tilde{w}_i)}$ , we have

$$\mathbb{E}_{\hat{Z}_i, U_i | \tilde{w}_i} \left[ g(\tilde{w}_i, \hat{Z}_i, U_i) \right] \leq \Delta_1(\tilde{w}_i) I(\hat{Z}_i; U_i | \widetilde{W}_i = \tilde{w}_i) + 0.72 \frac{\mathbb{E}_{\hat{Z}_i | \tilde{w}_i} \left[ \left( \ell(\tilde{w}_i^-, \hat{Z}_i) - \ell(\tilde{w}_i^+, \hat{Z}_i) \right)^2 \right]}{\Delta_1(\tilde{w}_i)}.$$

Let  $\Lambda(\tilde{w}_i) = \mathbb{E}_{\hat{Z}_i | \tilde{w}_i} \left[ \left( \ell(\tilde{w}_i^-, \hat{Z}_i) - \ell(\tilde{w}_i^+, \hat{Z}_i) \right)^2 \right] / \Delta_1(\tilde{w}_i)^2$ , then

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\widetilde{W}_i} \left[ \Delta_1(\widetilde{W}_i) \left( I^{\widetilde{W}_i}(\hat{Z}_i; U_i) + 0.72 \Lambda(\widetilde{W}_i) \right) \right].$$

For the second part, if  $\mathcal{A}$  is further  $\beta_2$ -uniform stable, recall Eq. (E.14) and by the non-decreasing property of  $h$ , we have

$$\mathbb{E}_{\hat{Z}_i, U_i | \tilde{w}_i} \left[ g(\tilde{w}_i, \hat{Z}_i, U_i) \right] \leq \inf_{t>0} \frac{I(\hat{Z}_i; U_i | \widetilde{W}_i = \tilde{w}_i)}{t} + h(\beta_2 t) t \mathbb{E}_{\hat{Z}_i | \tilde{w}_i} \left[ \left( \ell(\tilde{w}_i^-, \hat{Z}_i) - \ell(\tilde{w}_i^+, \hat{Z}_i) \right)^2 \right].$$

Let  $t = \frac{1}{\beta_2}$  and taking expectation over  $\widetilde{W}_i$ , we have

$$\begin{aligned} \mathbb{E}_{\hat{Z}_i, U_i, \widetilde{W}_i} \left[ g(\widetilde{W}_i, \hat{Z}_i, U_i) \right] &\leq \beta_2 I(\hat{Z}_i; U_i | \widetilde{W}_i) + 0.72 \frac{\mathbb{E}_{\hat{Z}_i, \widetilde{W}_i} \left[ \left( \ell(\widetilde{W}_i^-, \hat{Z}_i) - \ell(\widetilde{W}_i^+, \hat{Z}_i) \right)^2 \right]}{\beta_2} \\ &= \beta_2 I(\hat{Z}_i; U_i | \widetilde{W}_i) + 0.72 \frac{\gamma_4^2}{\beta_2}, \end{aligned}$$

where the equality is by the definition of  $\gamma_4$ -SCH-D.

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \beta_2 I(\hat{Z}_i; U_i | \widetilde{W}_i) + 0.72 \frac{\gamma_4^2}{\beta_2}.$$

This concludes the proof. □

#### E.4.4 Proof of Theorem 7.5.4

We present a stronger version of Theorem 7.5.4.

**Theorem E.4.1.** *Under the same conditions in Theorem 7.5.1, and we further assume that  $\mathcal{A}$  is  $\gamma_2$ -SCH-B stable and symmetric with respect to  $S$ , i.e. it does not depend on the order of the elements in the training sample. Let  $\bar{\Delta}_1(\widetilde{W}) = \frac{1}{n} \sum_{i=1}^n \Delta_1(\widetilde{W}_i)^2$ , we have*

$$\mathbb{E}_{W,S} [(L_S(W) - L_\mu(W))^2] \leq \frac{6}{n} \mathbb{E}_{\widetilde{W}} \left[ \bar{\Delta}_1(\widetilde{W}) \left( I^{\widetilde{W}}(E; U) + \frac{\log 3}{2} \right) \right] + \frac{1}{n} + 4\gamma_2^2.$$

Then Theorem 7.5.4 is a corollary of Theorem E.4.1.

*Proof of Theorem 7.5.4.* For  $\beta_2$ -uniform stable algorithm, by  $\bar{\Delta}_1(\widetilde{W}) \leq \beta_2^2$  and  $\gamma_2^2 \leq \beta_2^2$ , we have

$$\begin{aligned} \mathbb{E}_{W,S} [(L_S(W) - L_\mu(W))^2] &\leq \frac{6\beta_2^2}{n} \left( I(E; U|\widetilde{W}) + \frac{\log 3}{2} \right) + \frac{1}{n} + 4\beta_2^2 \\ &= 4\beta_2^2 \left( \frac{1.5I(E; U|\widetilde{W}) + 0.82}{n} + 1 \right) + \frac{1}{n}. \end{aligned}$$

This completes the proof.  $\square$

Before we prove Theorem E.4.1, we need to first obtain the following lemma.

**Lemma E.4.1.** *Under the same conditions in Theorem 7.5.1, let  $\bar{\Delta}_1(\widetilde{W}) = \frac{1}{n} \sum_{i=1}^n \Delta_1(\widetilde{W}_i)^2$ , we have*

$$\mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i)] - L_S(W) \right)^2 \right] \leq \frac{3}{n} \mathbb{E}_{\widetilde{W}} \left[ \bar{\Delta}_1(\widetilde{W}) \left( I^{\widetilde{W}}(E; U) + \frac{\log 3}{2} \right) \right].$$

*Proof of Lemma E.4.1.* Here we borrow some proof techniques used in [199, Eq. (2), Thm. 2].

Let  $g(\tilde{w}, \hat{z}_i, u_i) = (-1)^{u_i} (\ell(\tilde{w}_i^-, \hat{z}_i) - \ell(\tilde{w}_i^+, \hat{z}_i))$  and let  $G \sim \mathcal{N}(0, 1)$  be an independent



standard Gaussian random variable. Let  $f = t \cdot (\frac{1}{n} \sum_{i=1}^n g)^2$  in Lemma 2.4.1, then

$$\begin{aligned} & \mathbb{E}_{E,U|\tilde{w}} \left[ \left( \frac{1}{n} \sum_{i=1}^n g(\tilde{w}, \hat{Z}_i, U_i) \right)^2 \right] \\ & \leq \inf_{t>0} \frac{I(E; U | \widetilde{W} = \tilde{w}) + \log \mathbb{E}_{E,U'|\tilde{w}} \left[ e^{t(\frac{1}{n} \sum_{i=1}^n g(\tilde{w}, \hat{Z}_i, U'_i))^2} \right]}{t} \\ & = \inf_{t>0} \frac{I(E; U | \widetilde{W} = \tilde{w}) + \log \mathbb{E}_{E,U'|\tilde{w}} \left[ \mathbb{E}_G \left[ e^{\frac{G\sqrt{2t}}{n} \sum_{i=1}^n g(\tilde{w}, \hat{Z}_i, U'_i)} \right] \right]}{t} \end{aligned} \quad (\text{E.15})$$

$$\begin{aligned} & = \inf_{t>0} \frac{I(E; U | \widetilde{W} = \tilde{w}) + \log \mathbb{E}_{G,E|\tilde{w}} \left[ \prod_{i=1}^n \mathbb{E}_{U'_i} \left[ e^{\frac{G\sqrt{2t}}{n} g(\tilde{w}, \hat{Z}_i, U'_i)} \right] \right]}{t} \\ & \leq \inf_{t>0} \frac{I(E; U | \widetilde{W} = \tilde{w}) + \log \mathbb{E}_G \left[ e^{\frac{G^2 t \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}{n^2}} \right]}{t} \end{aligned} \quad (\text{E.16})$$

$$\begin{aligned} & \leq \inf_{t \in \left(0, \frac{n^2}{2 \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}\right)} \frac{I(E; U | \widetilde{W} = \tilde{w}) + \log \left( 1 / \sqrt{1 - \frac{2t \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}{n^2}} \right)}{t} \\ & = \inf_{t \in \left(0, \frac{n^2}{2 \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}\right)} \frac{I(E; U | \widetilde{W} = \tilde{w}) - \frac{1}{2} \log \left( 1 - \frac{2t \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}{n^2} \right)}{t}, \end{aligned} \quad (\text{E.17})$$

where Eq. (E.15) is by the moment generating function of Gaussian distribution:  $\mathbb{E}_G [e^{\lambda G}] = e^{\frac{\lambda^2}{2}}$  for all  $\lambda \in \mathbb{R}$ , Eq. (E.16) is by Lemma 2.4.2 and Eq. (E.17) is by the moment generating function of chi-squared distribution:  $\mathbb{E}_G [e^{\lambda G^2}] \leq \frac{1}{\sqrt{1-2\lambda}}$  for  $\lambda < \frac{1}{2}$ .

Let  $t = \frac{n^2}{3 \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}$  be substituted to the last equation above, we have

$$\mathbb{E}_{E,U|\tilde{w}} \left[ \left( \frac{1}{n} \sum_{i=1}^n g(\tilde{w}, \hat{Z}_i, U_i) \right)^2 \right] \leq \frac{3}{n^2} \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2 \left( I(E; U | \widetilde{W} = \tilde{w}) + \frac{\log 3}{2} \right). \quad (\text{E.18})$$

Let  $\bar{\Delta}_1(\tilde{w}) = \frac{1}{n} \sum_{i=1}^n \Delta_1(\tilde{w}_i)^2$ , and taking expectation over  $\widetilde{W}$  for both sides,

$$\mathbb{E}_{E,U,\widetilde{W}} \left[ \left( \frac{1}{n} \sum_{i=1}^n g(\widetilde{W}, \hat{Z}_i, U_i) \right)^2 \right] \leq \frac{3}{n} \mathbb{E}_{\widetilde{W}} \left[ \bar{\Delta}_1(\widetilde{W}) \left( I^{\widetilde{W}}(E; U) + \frac{\log 3}{2} \right) \right]. \quad (\text{E.19})$$

Applying Jensen's inequality to the square function, we have

$$\mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i)] - L_S(W) \right)^2 \right] \leq \mathbb{E}_{E,U,\widetilde{W}} \left[ \left( \frac{1}{n} \sum_{i=1}^n g(\widetilde{W}, \widehat{Z}_i, U_i) \right)^2 \right].$$

Combining Eq. (E.19) with the inequality above will concludes the proof.  $\square$

*Proof of Theorem E.4.1.* We are now in a position to prove Theorem E.4.1.

$$\begin{aligned} & \mathbb{E}_{W,S} [(L_S(W) - L_\mu(W))^2] \\ = & \mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \ell(W, Z_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i)] + \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i)] - \mathbb{E}_{Z'} [\ell(W, Z')] \right)^2 \right] \\ \leq & \underbrace{2 \mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \ell(W, Z_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i)] \right)^2 \right]}_{B_1} \\ & + 2 \underbrace{\mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i)] - \mathbb{E}_{Z'} [\ell(W, Z')] \right)^2 \right]}_{B_2}. \end{aligned}$$

Notice that  $B_1$  can be bounded by using Lemma E.4.1. We now focus on  $B_2$ :

$$\begin{aligned} & \mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i)] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Z'} [\ell(W, Z')] - \mathbb{E}_{W^i|W} [\ell(W^i, Z')] + \mathbb{E}_{W^i|W} [\ell(W^i, Z')] \right)^2 \right] \\ = & \mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i) - \mathbb{E}_{Z'} [\ell(W^i, Z')]] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Z'} [\ell(W, Z') - \mathbb{E}_{W^i|W} [\ell(W^i, Z')]] \right)^2 \right] \\ \leq & \underbrace{2 \mathbb{E}_{W,S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W^i|W} [\ell(W^i, Z_i) - \mathbb{E}_{Z'} [\ell(W^i, Z')]] \right)^2 \right]}_{B_3} \\ & + 2 \underbrace{\mathbb{E}_W \left[ \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Z'} [\ell(W, Z') - \mathbb{E}_{W^i|W} [\ell(W^i, Z')]] \right)^2 \right]}_{B_4}. \end{aligned}$$

For  $B_3$ , we apply Jensen's inequality to move the expectation over  $W^i$  outside of the square function,

$$\begin{aligned} B_3 &\leq \mathbb{E}_{W^1, W^2, \dots, W^n, S} \left[ \left( \frac{1}{n} \sum_{i=1}^n \ell(W^i, Z_i) - \mathbb{E}_{Z'_i} [\ell(W^i, Z')] \right)^2 \right] \\ &= \mathbb{E}_{S', S, R} \left[ \left( \frac{1}{n} \sum_{i=1}^n \ell(\mathcal{A}(S^i, R), Z_i) - \mathbb{E}_{Z'} [\ell(\mathcal{A}(S^i, R), Z')] \right)^2 \right]. \end{aligned}$$

Notice that  $S'$ ,  $S$  and  $R$  are all independent with each other (so  $W^i$ ,  $Z_i$  and  $Z'_i$  are also independent with each other). If we further let  $\mathcal{A}$  be symmetric, namely  $W^i$  does not dependent on  $i$ , then the inequality above is equivalent to

$$\begin{aligned} B_3 &\leq \mathbb{E}_{W, S'} \left[ \left( \frac{1}{n} \sum_{i=1}^n \ell(W, Z'_i) - \mathbb{E}_{Z'} [\ell(W, Z')] \right)^2 \right] \\ &= \mathbb{E}_W \left[ \mathbb{E}_{S'} \left[ \left( \frac{1}{n} \sum_{i=1}^n \ell(W, Z'_i) - \mathbb{E}_{Z'} [\ell(W, Z')] \right)^2 \right] \right]. \end{aligned}$$

Hence, the inner expectation in the RHS above is just the variance of the sample mean of  $n$  i.i.d bounded random variables. Recall that  $\ell(\cdot, \cdot) \in [0, 1]$ , thereby

$$B_3 \leq \mathbb{E}_W \left[ \frac{\text{Var}_{Z'}(\ell(W, Z'))}{n} \right] \leq \frac{1}{4n},$$

where the second inequality is by Lemma 2.4.3.

Then, for  $B_4$ , we also apply Jensen's inequality to the square function, and by the definition of  $\gamma_2$ -SCH-B stability, we have

$$\begin{aligned} B_4 &\leq \mathbb{E}_{W, Z'} \left[ \left( \frac{1}{n} \sum_{i=1}^n \ell(W, Z') - \mathbb{E}_{W^i|W} [\ell(W^i, Z')] \right)^2 \right] \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W, Z'} \left[ (\ell(W, Z') - \mathbb{E}_{W^i|W} [\ell(W^i, Z')])^2 \right] \leq \gamma_2^2. \end{aligned}$$

Putting everything together, we have

$$\begin{aligned}
\mathbb{E}_{W,S} [(L_S(W) - L_\mu(W))^2] &\leq 2B_1 + 2B_2 \\
&\leq 2B_1 + 4B_3 + 4B_4 \\
&\leq 2B_1 + \frac{1}{n} + 4\gamma_2 \\
&\leq \frac{6}{n} \mathbb{E}_{\widetilde{W}} \left[ \bar{\Delta}_1(\widetilde{W}) \left( I^{\widetilde{W}}(E; U) + \frac{\log 3}{2} \right) \right] + \frac{1}{n} + 4\gamma_2^2,
\end{aligned}$$

where the last inequality is by Lemma E.4.1. This completes the proof.  $\square$

#### E.4.5 Proof of Theorem 7.5.5

*Proof.* Let  $g(\tilde{z}_i^+, w_i, \bar{w}_i, u_i) = (-1)^{u_i} (\ell(\bar{w}_i, \tilde{z}_i^+) - \ell(w_i, \tilde{z}_i^+))$ . Again, by Lemma 2.4.1, we have

$$\mathbb{E}_{W_i, \bar{W}_i, U_i | \tilde{z}_i^+} [g(\tilde{z}_i^+, W_i, \bar{W}_i, U_i)] \leq \inf_{t > 0} \frac{I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+ = \tilde{z}_i^+) + \log \mathbb{E}_{W_i, \bar{W}_i, U_i' | \tilde{z}_i^+} [e^{tg(\tilde{z}_i^+, W_i, \bar{W}_i, U_i')}] }{t}. \quad (\text{E.20})$$

Similar to the previous proofs, it's easy to see that  $g(\tilde{z}_i^+, W_i, \bar{W}_i, U_i)$  is a zero-mean random variable bounded in  $[-\Delta_2(\tilde{z}_i^+), \Delta_2(\tilde{z}_i^+)]$ . Thus,

$$\mathbb{E}_{W_i, \bar{W}_i, U_i | \tilde{z}_i^+} [g(\tilde{z}_i^+, W_i, \bar{W}_i, U_i)] \leq \inf_{t > 0} \frac{I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+ = \tilde{z}_i^+) + \frac{t^2 \Delta_2(\tilde{z}_i^+)^2}{2}}{t}. \quad (\text{E.21})$$

To prove the first bound, we let  $t = \sqrt{\frac{I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+ = \tilde{z}_i^+)}{2\Delta_2(\tilde{z}_i^+)^2}}$ , then

$$\mathbb{E}_{W_i, \bar{W}_i, U_i | \tilde{z}_i^+} [g(\tilde{z}_i^+, W_i, \bar{W}_i, U_i)] \leq \Delta_2(\tilde{z}_i^+) \sqrt{2I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+ = \tilde{z}_i^+)}.$$

Recall Eq. (E.3) in Lemma E.1.1, hence,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tilde{Z}_i^+} \left[ \Delta_2(\tilde{Z}_i^+) \sqrt{2I^{\tilde{Z}_i^+}(W_i, \bar{W}_i; U_i)} \right]. \quad (\text{E.22})$$

To prove the second bound, we take expectation over  $\tilde{Z}_i^+$  for Eq. (E.21),

$$\mathbb{E}_{W_i, \bar{W}_i, U_i, \tilde{Z}_i^+} \left[ g(\tilde{Z}_i^+, W_i, \bar{W}_i, U_i) \right] \leq \inf_{t>0} \frac{I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+) + \frac{t^2 \mathbb{E}_{\tilde{Z}_i^+} [\Delta_2(\tilde{Z}_i^+)^2]}{2}}{t}.$$

Let  $t = \sqrt{\frac{I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)}{2 \mathbb{E}_{\tilde{Z}_i^+} [\Delta_2(\tilde{Z}_i^+)^2]}}$ , then

$$\mathbb{E}_{W_i, \bar{W}_i, U_i, \tilde{Z}_i^+} \left[ g(\tilde{Z}_i^+, W_i, \bar{W}_i, U_i) \right] \leq \sqrt{2 \mathbb{E}_{\tilde{Z}_i^+} [\Delta_2(\tilde{Z}_i^+)^2] I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)}.$$

Ergo,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2 \mathbb{E}_{\tilde{Z}_i^+} [\Delta_2(\tilde{Z}_i^+)^2] I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+)}. \quad (\text{E.23})$$

This completes the proof.  $\square$

## E.5 Omitted Proof in Section 7.7

### E.5.1 Proof of Proposition 7.7.1

*Proof.* By Jensen's inequality and triangle inequality, for any  $i \in [n]$ , we have

$$\begin{aligned} & \mathbb{E}_{S,R,Z'} \left[ \left( \ell(W, Z') - \mathbb{E}_{W^i|W} [\ell(W^i, Z')] \right)^2 \right] \\ & \leq \mathbb{E}_{W,W^i,Z'} \left[ \left( \ell(W, Z') - \ell(W^i, Z') \right)^2 \right] \\ & = \mathbb{E}_{W,W^i,Z'} \left[ \left( \ell(W, Z') - \ell(w^*, Z') + \ell(w^*, Z') - \ell(W^i, Z') \right)^2 \right] \\ & \leq 2 \mathbb{E}_{W,Z'} \left[ \left( \ell(W, Z') - \ell(w^*, Z') \right)^2 \right] + 2 \mathbb{E}_{W^i,Z'} \left[ \left( \ell(W^i, Z') - \ell(w^*, Z') \right)^2 \right] \\ & \leq 4B \mathbb{E}_W \left[ \left( L_\mu(W) - L_\mu(w^*) \right)^{\frac{1}{\kappa}} \right], \end{aligned}$$

where the last inequality is by the definition of the Bernstein condition. This completes the proof.  $\square$

### E.5.2 Proof of Theorem 7.7.1

*Proof.* Let  $g(\Delta\ell_i, u_i) = (-1)^{u_i} \Delta\ell_i$ . Notice that  $|g(\Delta L_i, U'_i)| \leq \beta_2$  and  $g(\Delta L_i, U'_i)$  is again zero-mean, then

$$\begin{aligned} \mathbb{E}_{\Delta L_i, U_i} [g(\Delta L_i, U_i)] &\leq \frac{I(\Delta L_i; U_i) + \log \mathbb{E}_{\Delta L_i, U'_i} [e^{tg(\Delta L_i, U'_i)}]}{t} \\ &\leq \beta_2 \sqrt{2I(\Delta L_i; U_i)}. \end{aligned}$$

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\beta_2}{n} \sum_{i=1}^n \sqrt{2I(\Delta L_i; U_i)}.$$

To prove the disintegrated CMI bound, we let  $g$  be defined in the same way, and the remaining development is the same with the proof in Theorem 7.5.5.

For the second inequality, notice that  $I(\Delta L_i; U_i) \leq I(\Delta L_i; U_i | \tilde{Z}_i^+)$  by using the chain rule of mutual information and the independence between  $\tilde{Z}_i^+$  and  $U_i$ . In addition, moving the expectation over  $\tilde{Z}_i^+$  inside the square-root function by Jensen's inequality, we have  $\mathbb{E}_{\tilde{Z}_i^+} \sqrt{I(\tilde{Z}_i^+; \Delta L_i; U_i)} \leq \sqrt{I(\Delta L_i; U_i | \tilde{Z}_i^+)}$ .  $\square$

## E.6 CLB Examples

In Example 7.1, [78, Thm. 17] demonstrates the non-vanishing behavior of individual IOMI and e-CMI. This is primarily attributed to the dimension-dependent nature of IOMI and CMI. Specifically, there are certain dimensional settings where IOMI can grow faster than  $\mathcal{O}(n)$ , as shown in [78, Thm.4], and CMI approaches a certain fraction of its upper bound, as illustrated in Example 7.1, resulting in non-vanishing behavior. Specifically, in Example 7.1, [78] employs the birthday paradox [144, Sec. 5.1] problem to demonstrate that for a large value of  $d$ , the probability that no pair of instances in  $\tilde{Z}$  sharing the same non-zero coordinate (referred to as event  $E_0$ ) is smaller than a constant probability (that could be independent of  $n$ ). Particularly, it is shown that if  $d \geq \frac{2n-1}{1-c^{1/(2n-1)}}$ , then  $P(E_0) \geq c \geq (1 - \frac{2n-1}{d})^{2n-1}$ . As an example, [78] chooses  $d = 2n^2$ , resulting in  $c \geq 0.1$ .

**Failure of  $I(W; Z_i)$**  Consider the case where  $d = 2n^2$ . For the individual CMI [179, 247],  $I(W; U_i | \tilde{Z}_i)$ , we have the following:

$$I(W; U_i | \tilde{Z}_i) = \log 2 - H(U_i | W, \tilde{Z}_i) \geq 0.1 \cdot \log 2.$$

This inequality holds because when event  $E_0$  does not occur, one can determine the value of  $U_i$  completely, as the returned hypothesis is a weighted sum of the sample. In other words, examining the non-zero coordinates of  $W$  is sufficient to determine  $U_i$ . For an in-depth derivation of this inequality, readers are referred to the updated version of [78], where their corrected proof involves Fano’s inequality. Furthermore, using the relation  $I(W; Z_i) \geq I(W; U_i | \widetilde{Z}_i)$  [247], we conclude that  $I(W; Z_i) \in \Omega(1)$ .

**Failure of  $I(\widehat{Z}_i; U_i | \widetilde{W}_i)$**  Notably, our hypotheses-conditioned CMI also does not vanish for the same reason. More precisely, when  $\widetilde{W}_i$  and  $\widehat{Z}_i$  are given, there exists a constant probability (independent of  $n$ ) that allows us to fully determine the returned hypothesis based on  $\widehat{Z}_i$ , thereby determining the value of  $U_i$ .

**Failure of  $I(\Delta L_i; U_i)$**  Furthermore, even the loss difference based CMI (e.g., as shown in Theorem 7.7.1), which provides the tightest CMI measure, still does not vanish. This is attributed to the fact that if the hypothesis  $W$  is independent of certain  $Z$ , there exists a constant probability where the loss becomes zero (recall that the loss is the negative inner product of  $W$  and  $Z$ ). Consequently, one can determine the value of  $U_i$  by observing the sign of the random variable  $\Delta L_i$ . This also indicates the limitations of e-CMI and  $f$ -CMI in capturing the generalization behavior for Example 7.1.

In Example 7.2, following the approach in [78], the training sample  $S = \{Z_i\}_{i=1}^n \sim \mu^n$  can be represented as  $S = \frac{z_0}{R_0} (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ , where  $\{\varepsilon_i\}_{i=1}^n$  is a sequence of independent Rademacher random variables, i.e.,  $\varepsilon_i \sim \text{Unif}(\{-1, 1\})$ . The empirical risk is given by  $L_S(W) = -\frac{L}{nR_0} \langle W, \sum_{i=1}^n \varepsilon_i \rangle$ . In this case, the ERM solution is  $W_{ERM} = z_0$  if  $\text{sign}(\sum_{i=1}^n \varepsilon_i) = 1$ , and  $W_{ERM} = -z_0$  if  $\text{sign}(\sum_{i=1}^n \varepsilon_i) = -1$ . It is clear that

$$\sup_{w, w^i, z} |\ell(w, z) - \ell(w^i, z)| \leq \sup_{w, w^i, z} L \|w - w^i\| \leq 2LR_0.$$

Hence, we observe that  $\beta_2$  is now a constant, whereas IOMI has an upper bound:  $\sum_{i=1}^n I(W; Z_i) \leq I(W; S) \leq I(W; \text{sign}(\sum_{i=1}^n \varepsilon_i)) \leq H(\text{sign}(\sum_{i=1}^n \varepsilon_i)) \leq 1$ , where the second inequality follows from the Markov chain  $S - \text{sign}(\sum_{i=1}^n \varepsilon_i) - W$ . This provides us with a generalization bound of  $\frac{2LR_0}{\sqrt{n}}$ . Meanwhile, the actual generalization error satisfies  $\mathcal{E}_\mu(\mathcal{A}) \geq \frac{LR_0}{\sqrt{2n}}$  (see [78, AppendixB] for a derivation). Thus, the IOMI bound is tight up to a constant, and the stability bound  $\beta_2$  itself is vacuous. It is worth noting that  $I(\widehat{Z}_i; U_i | \widetilde{W}_i) \leq I(W; Z_i) \leq 1$  by Theorem 7.5.2, indicating that the CMI bound is also tight.

We would like to note that the failures of chained mutual information bounds [11] are not demonstrated in the counterexamples presented in [78]. Notably, when the hypothesis is quantized, it becomes more challenging to guess  $U_i$  or  $Z_i$ . Therefore, exploring the potential of chained information-theoretic bounds, which do not necessarily rely on stability notions, could be another avenue to explain the generalization behavior observed in these counterexamples.

## E.7 Additional Applications

### E.7.1 Compression Schemes

We now consider the algorithm that has a compression scheme [127]. Formally, a sample compression scheme of size  $k \in \mathbb{N}$  is a pair of maps  $(\mathcal{A}_1, \mathcal{A}_2)$ . Specifically, for all samples  $s$  with  $n > k$ ,  $\mathcal{A}_1 : \mathcal{Z}^n \rightarrow \mathcal{Z}^k$  compresses the sample into a length- $k$  subsequence  $\mathcal{A}_1(s) \subseteq s$ . Then  $\mathcal{A}_2 : \mathcal{Z}^k \rightarrow \mathcal{W}$  could be some arbitrary mapping. Hence,  $\mathcal{A}(\cdot) = \mathcal{A}_2(\mathcal{A}_1(\cdot))$ . Let  $K$  be the index set for  $S$  selected by  $\mathcal{A}_1$ , and let  $\bar{K}$  be the selected index set for  $S^i$ . In this case, our supsample-conditioned CMI has an upper bound:  $I(W_i, \bar{W}_i; U_i | \tilde{Z}_i^+) \leq I(K, \bar{K}; U_i | \tilde{Z}_i^+) \leq H(K, \bar{K} | \tilde{Z}_i^+) \leq 2 \log \binom{n}{k} \leq 2k \log n$ . Then, if  $\mathcal{A}$  is further  $\beta_2$ -uniform stable, then we have the generalization bound  $\mathcal{E}_\mu(\mathcal{A}) \leq \mathcal{O}(\beta_2 \sqrt{k \log n})$ . If  $\beta_2 < \mathcal{O}(1/\sqrt{n})$ , this bound improves the bound in [199]. It is unclear if we can obtain any improved bound for *stable* compression schemes [25], in which case [75] provides an optimal bound that removing the  $\log n$  factor for the realizable setting. A main difficulty is that an interpolating algorithm is usually unstable due to the fitting-stability tradeoff [186, Sec. 13.4].

### E.7.2 Distillation Algorithm

The high-probability generalization property of distillation algorithm is studied in [57]. In the first training stage of distillation, we obtain a  $w_s^*$  from a highly complex hypothesis space  $\mathcal{W}_1$  based on a training sample  $s$ . Same to [57], we assume that the first learning stage is  $\alpha$ -sensitive, namely  $\|w_s^* - w_{s^i}^*\| \leq \alpha = \mathcal{O}(1/n)$ . In the second stage, the algorithm  $\mathcal{A}$  will select a hypothesis that is  $\lambda$ -close to  $w_s^*$  from a less complex hypothesis space  $\mathcal{W}_2 = \{w \in \mathcal{W} : \|w - w_s^*\|_\infty \leq \lambda\}$ . Let the loss function  $\ell$  be  $L$ -Lipschitz with respect to the first argument. Consequently,  $\gamma_3 \leq L\|w_s^* - w_{s^i}^*\| \leq L\alpha$ . Then, by Theorem 7.5.1, we have  $\mathcal{E}_\mu(\mathcal{A}) \leq L\alpha \frac{1}{n} \sum_{i=1}^n \sqrt{2H(U_i)} = \sqrt{2 \log 2} L\alpha$ . Notice that the loss here may not necessarily be bounded or subgaussian, rendering previous bounds inapplicable.



### E.7.3 Regularized Empirical Risk Minimization

Regularized Empirical Risk Minimization (ERM) learning rules involve minimizing the empirical risk and a regularization function jointly:  $\arg \min_{w \in \mathcal{W}} L_S(w) + f_{\text{reg}}(w)$ , where  $f_{\text{reg}} : \mathcal{W} \rightarrow \mathbb{R}$ . Here we specifically consider Tikhonov regularization [186], namely  $f_{\text{reg}}(w) = \lambda \|w\|^2$ , where  $\lambda > 0$  is a tradeoff coefficient. The regularized ERM algorithm  $\mathcal{A}$  aims to find

$$w = \arg \min_{w \in \mathcal{W}} L_S(w) + \lambda \|w\|^2.$$

This regularization term ensures strong convexity of the training objective. Based on Theorem 7.5.3, we can derive the following results.

**Corollary E.7.1.** *Assume that the loss function  $\ell$  is convex and  $L$ -Lipschitz. Then, for the regularized ERM algorithm with Tikhonov regularization, we have*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{2L^2}{\lambda n} \left( \frac{1}{n} \sum_{i=1}^n I(\hat{Z}; U_i | \widetilde{W}_i) + 0.72 \right).$$

*Proof of Corollary E.7.1.* By invoking [186, Corollary 13.6], we know that  $\gamma_4 \leq \beta_2 = \frac{2L^2}{\lambda n}$ . Plugging the value of  $\beta_2$  will give us the desired result.  $\square$

**Corollary E.7.2.** *Assume that the loss function is  $\rho$ -smooth and nonnegative. Let  $\lambda \geq \frac{2\rho}{n}$ . Then, for the regularized ERM algorithm with Tikhonov regularization, we have*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{48\rho\hat{L}_n}{\lambda n} \left( \frac{1}{n} \sum_{i=1}^n I(\hat{Z}; U_i | \widetilde{W}_i) + 0.72 \right).$$

*Proof of Corollary E.7.2.* By invoking [186, Corollary 13.7], we know that  $\gamma_4 \leq \beta_2 = \frac{48\rho\hat{L}_n}{\lambda n}$ . Plugging the value of  $\beta_2$  will give us the desired result.  $\square$

Although these bounds do not enhance the convergence rate of  $\mathcal{O}(1/n)$  in these settings, they consistently offer tighter results compared to uniform stability-based bounds if  $\frac{1}{n} \sum_{i=1}^n I(\hat{Z}; U_i | \widetilde{W}_i) \leq 0.28$ . In addition, the expected empirical risk  $\hat{L}_n$  appears in the bound of Corollary E.7.2. While  $\lambda$  has a lower bound,  $\hat{L}_n$  could not be arbitrarily small for the regularized ERM.

Notice that previous information-theoretic bounds could not obtain the convergence rate of  $\mathcal{O}(1/n)$  as in our results unless ICMI or CMI itself decays with  $\mathcal{O}(1/n)$ .