

INFORMATION-THEORETIC MEASURES OF GENERALIZATION
IN MACHINE LEARNING

by

Mahdi Haghifam

A thesis submitted in conformity with the requirements
for Doctor of Philosophy

Electrical & Computer Engineering
University of Toronto

© Copyright 2023 by Mahdi Haghifam

Information-Theoretic Measures of Generalization
in Machine Learning

Mahdi Haghifam
Doctor of Philosophy

Electrical & Computer Engineering
University of Toronto
2023

Abstract

Machine learning (ML) technologies are rapidly being adopted across various industries, scientific fields, and governments. However, despite the widespread use of ML, the research community still lacks a thorough understanding of the principles that explain when and why these technologies work in real-world applications.

One of the primary challenges to deploying ML methods is ensuring generalization, or the ability of the model to perform well on unseen data as well as on training data. In this thesis, we explore generalization from an information-theoretic perspective. This perspective allows us to examine the impact of learning algorithms and data distribution on generalization error, a crucial feature that offers an opportunity to address an essential open challenge in generalization theory.

Our main goals in this thesis are twofold: (I) to explain the performance of existing learning algorithms used to train modern deep neural networks and (II) to unify and simplify the landscape of generalization frameworks using the language of information theory.

For goal (I), we propose a novel approach based on data-dependent estimation to estimate the mutual information terms that appear in generalization bounds. We demonstrate the applicability of this technique in studying the generalization error of Stochastic Gradient Langevin Dynamics (SGLD). Our numerical analyses show that the obtained bounds are the first non-vacuous generalization guarantees for SGLD and its full-batch counterpart for modern deep-learning datasets and architectures.

Regarding goal (II), we explore the connections between information-theoretic frameworks for generalization and classical min-max frameworks for generalization. Our results indicate that information-theoretic frameworks are highly expressive in the context of binary classification. However, we also uncover some limitations of information-theoretic techniques when analyzing stochastic gradient descent in the context of stochastic convex optimization.

To Mom, Dad,
and Shiva.

Acknowledgements

Admittedly, the journey to this thesis has not been a paved way; it has gone through difficult paths, and it has been full of ups and downs. However, during these years I have been extremely lucky to have been surrounded by wonderful people. This is an opportunity to recognize my appreciation to them.

First, I would like to express my gratitude to my advisor Daniel Roy. I first met him in November 2018. I was a naive student with a little background in Machine Learning and also little confidence. The meeting lasted for an hour, Dan was the only speaker, and I could make sense out of nothing. I only remember his brightness and enthusiasm from the meeting. He took a chance on me in 2018 and opened many doors for me. He has given me unparalleled opportunity for learning, thinking, and more importantly self-discovery. He is kind, supportive, respectful, and always available to help by any means he could, both academically and non-academically. His mentorship has shaped the way I read, the way I write, the way I speak, and the way I think. I will forever be indebted to Dan for his guidance and unwavering belief in my potential.

Next, I would like to thank Gintare Karolina Dziugaite. My journey in Machine learning literally started by an email to Karolina to discuss the possibility of working with her as an intern. She gave me the opportunity and it truly changed my career path. This marked the beginning of an exceptionally productive and transformative period in my life. Moreover, Karolina has been an unwavering source of support in navigating both personal and professional challenges during my PhD. I am forever indebted to her for her invaluable guidance and support.

I would like to thank all my collaborators: Jeffery Negrea, Shay Moran, Borja Rodriguez Galvez, Gregley Neu, Ashish Khisti, and Vincent Tan for everything they have taught me. Specifically, I want to thank Jeffery Negrea, my first collaborator. I was fortunate and lucky to work with him during my first years. I would also like to thank Blair Bilodeau, Mufan Li, Yasaman Mahdaviyeh, and Ekansh Sharma.

I spent Summer 2022 working with Thomas Steinke and Abhradeep Guha Thakurta as an intern at Google brain. Thomas and Abhradeep introduced me to the wonderful area of privacy. Their unique problem-solving skills, their ability to pinpoint important research questions, and their clarity of thought have been a continuous source of inspiration for me. I hope I can follow their approach as a researcher, and I am immensely grateful for the knowledge and skills I gained while working alongside them.

I would like to thank my parents, Mahmoudreza Haghifam and Parvin Vadoudi-Mofid, for their unconditional love and support throughout my entire life. My parents also imparted in me a profound love for science that has been a driving force in my

life. Their continuous encouragement and prayer is what has motivated me in my life, and I am forever indebted to them for their sacrifices. You are the best family I could ever dream of. I am especially grateful that my lovely sister has joined me here in Toronto. Because of her, I feel more complete, as a big part of me is now closer to me.

Lastly, but by no means least, I want to express my heartfelt gratitude to my best friend and the love of my life, Shiva Ketabi. Words alone cannot adequately convey my appreciation for your unwavering love, the cherished memories we've created together, and the emotional support you've provided during the toughest days of my journey. You have been an integral part of all I have accomplished, and I am thankful for your presence in my life. I am incredibly excited to see what we can achieve together in the future!

Attribution

The content of this thesis covers several papers developed in collaboration with other researchers, and the results herein were possible only through the effort and insight of everyone involved.

1. The content of Chapter 2 is based on:

- Negrea*, J., Haghifam*, M., Dziugaite, G. K., Khisti, A., Roy, D. M. (2019). Information-theoretic generalization bounds for SGLD via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32.

Jeffery Negrea and I are equal-contribution authors of this work, and the order is determined by a coin flip.

2. The content of Chapter 3 is based on the following work:

- Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., Dziugaite, G. K. (2020). Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33, 9925-9935.

3. The content of Chapter 4 is based on:

- Haghifam, M., Dziugaite, G. K., Moran, S., Roy, D. (2021). Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34, 26370-26381.

4. The content of Chapter 5 is based on:

- Haghifam, M., Moran, S., Roy, D. M., Dziugaite, G. K. (2022). Understanding Generalization via Leave-One-Out Conditional Mutual Information. In *2022 IEEE International Symposium on Information Theory (ISIT)* (pp. 2487-2492). IEEE.

5. The content of Chapter 6 is based on:

- Haghifam*, M., Rodriguez-Galvez*, B., Thobaben, R., Skoglund, M., Roy, D. M., Dziugaite, G. K. (2023). Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In *International Conference on Algorithmic Learning Theory* (pp. 663-706). PMLR.

Borja Rodriguez-Galvez and I are equal-contribution authors of this work.

- Neu, G., Dziugaite, G. K., Haghifam, M., Roy, D. M. (2021). Information-theoretic generalization bounds for stochastic gradient descent. In Conference on Learning Theory (pp. 3526-3545). PMLR.

I was a co-author of this work.

Contents

1	Introduction	1
1.1	What is Learning from Data?	1
1.2	Information-Theoretic Measures for Generalization	2
1.3	Main Questions and Overview of the Results	4
1.3.1	Characterizing Generalization Error of Iterative Learning Algorithms	4
1.3.2	Unifying Framework for Generalization	7
2	Generalization Bounds for SGLD via Data-Dependent Estimates	10
2.1	Introduction	10
2.1.1	Preliminaries	12
2.2	Methods	12
2.2.1	Information-Theoretic Generalization Bounds based on Random Subsets of Data	14
2.2.2	Decomposing KL Divergences and Mutual Information for Sequential Algorithms	16
2.3	Generalization Bounds for Specific Algorithms	17
2.3.1	Stochastic Gradient Langevin Dynamics	18
2.3.2	Langevin Dynamics	20
2.4	Empirical Results	21
3	Generalization Bounds based on Conditional Mutual Information	24
3.1	Introduction	24
3.1.1	Contributions	26
3.1.2	Definitions from Probability and Information Theory	26
3.2	Connections between $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$	27
3.3	Sharpened Bounds based on Individual Samples	28
3.3.1	Controlling CMI bounds using KL Divergence	30

3.3.2	Tighter Generalization bound for the case $m = 1$	30
3.4	Generalization bounds for noisy, iterative algorithms	31
3.4.1	Bounding Generalization Error via Hypothesis Testing	31
3.4.2	Example: Langevin Dynamics Algorithm for Non-Convex Learning	32
4	Unified Information-Theoretic Framework for Generalization	39
4.1	Introduction	39
4.1.1	Contributions	41
4.2	Preliminaries	42
4.2.1	Conditional mutual information (CMI) of an algorithm	43
4.2.2	Measures of divergence and information	43
4.3	Optimal CMI Bound for SVM and Stable Compression Schemes	45
4.3.1	CMI of Stable Compression Schemes	46
4.4	CMI of Proper Learning of VC classes	49
4.4.1	A Limitation of Proper Learning	49
4.4.2	VC Classes with Finite Star Number	50
4.5	Universality of eCMI and Improper Learning of VC Classes	53
4.5.1	The One-Inclusion Graph Prediction Strategy	54
5	Leave-One-Out Conditional Mutual Information	56
5.1	Introduction	56
5.1.1	Related Work	58
5.1.2	Notation	58
5.2	Generalization Bounds	59
5.3	A Connection with leave-one-out error	61
5.3.1	Universality of LOO^eCMI	64
5.4	An Optimal Bound for Learning VC classes	65
5.5	A Hierarchy of Measures of Information	67
6	Limitations of Information-Theoretic Generalization Bounds	72
6.1	Introduction	72
6.1.1	Contributions	74
6.1.2	Related Work	75
6.2	Preliminaries	76
6.2.1	Probability and Information Theory Notation	76
6.2.2	Stochastic Convex Optimization	76
6.2.3	Excess Risk of Gradient Descent	78

6.3	Main Questions and Overview of the Results	79
6.4	Information-Theoretic Generalization Bounds for the CLB setting	83
6.5	Failure of Information-Theoretic Bounds for GD in the CLB Setting	84
6.6	Implications for PAC-Bayes Bounds	87
6.7	Failure of Information-Theoretic Alternatives to the IOMI and CMI Frameworks	88
6.7.1	Information-Theoretic Alternatives to the IOMI and CMI Frameworks	89
6.7.2	Failure of the Alternatives	91
7	Discussion and Future Directions	93
7.1	Future Directions for Chapter 2 and Chapter 3	94
7.2	Future Directions for Chapter 4	94
7.3	Future Directions for Chapter 5	95
7.4	Future Directions for Chapter 6	96
A	Appendix of Chapter 2	98
A.1	Common Definitions	98
A.2	Proofs of Results	99
A.2.1	Bounding Mutual Information by KL Divergence	99
A.2.2	Proofs of Main Results	99
A.3	Mutual Information Bound for Subgaussian Losses	103
A.4	Properties of the Hypergeometric Distribution and of Finite Population Variances	105
A.4.1	Properties of the Hypergeometric Distribution	105
A.4.2	Finite Population Statistics with Disjoint Samples	106
A.5	Asymptotic Results	109
A.5.1	Langevin Dynamics	109
A.6	Comparing Theorems 2.2.3 to 2.2.5 when $m = n - 1$	110
A.7	An analytically tractable example	111
A.8	Experiment Details	112
A.8.1	Evaluation of the generalization bound	112
A.8.2	Learning Rate and Inverse Temperatures for Figs. 2.1b and 2.1c	113
A.8.3	Hyperparameters of our experiments	113
A.9	High Probability PAC-Bayes Bounds	114

B Appendix of Chapter 3	117
B.1 CMI, Membership Attack, and Fano’s Inequality	117
B.2 Matching the leading coefficient of Theorem 3.1.1 with $\text{CMI}_{\mathcal{D}}(\mathcal{A})$. . .	118
B.3 Proofs of Section 3.2	119
B.4 Proofs of Section 6.4	124
B.5 Proofs of Section 3.4	129
B.6 Conditional Han’s Inequality	132
B.7 Details of Experiments	132
B.7.1 Network architectures and learning curve	133
B.7.2 Optimizing the bound over the choice of θ function	135
C Appendix of Chapter 4	136
C.1 Known Bounds for Learning VC Classes	136
C.2 Proof of Theorem 4.2.3	137
C.3 Proof of Theorem 4.4.4	137
C.4 Proof of Theorem 4.4.6	138
C.5 Proof of Theorem 4.4.8	140
C.6 Proof of Theorem 4.4.9	143
C.7 Proof of Theorem 4.5.1	145
C.8 Proof of Theorem 4.5.6	146
C.8.1 Proof of Theorem 4.5.6	147
C.8.2 Proof of Theorem C.8.1	149
C.9 Description of the One-Inclusion Graph Prediction Algorithm	152
D Appendix of Chapter 6	154
D.1 Proof of the information-theoretic bounds of $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ in the CLB setting	154
D.1.1 Proof of Theorem 6.7.1: Individual-sample IOMI	154
D.1.2 Proof of Theorem 6.7.1: Individual-sample CMI	155
D.1.3 Proof of Theorem 6.7.4	155
D.2 Proof of Theorem 6.4.2	156
D.3 Proof of Theorem 6.5.1	157
D.3.1 Construction	158
D.3.2 Dynamics of GD	158
D.3.3 Noise with Large Variance Fails	162
D.3.4 Noise With Small Variance Fails: IOMI	167
D.3.5 Noise with Small Variance Fails: CMI	172

D.4	Proof of Theorem 6.6.2	176
D.4.1	Lower Bound on the Residual	176
D.4.2	Lower Bound on the Conditional-PAC Bayes Bound	176
D.4.3	Lower Bound on the Classical PAC-Bayes Bound	177
D.5	Proof of Theorem 6.7.7	183
D.5.1	Individual conditional mutual information	184
D.5.2	Evaluated conditional mutual information	186
D.6	Helper Lemmata	187
Bibliography		190

List of Tables

3.1	Summary of the results. The generalization bounds are reported at the end of training.	38
A.1	Details of Experiments reported in Fig. 2.1a for MNIST with MLP . . .	113
A.2	Details of Experiments reported in Figs. 2.1b to 2.1d for MNIST with CNN	114
A.3	Details of Experiments reported in Fig. 2.1e for Fashion-MNIST	114
A.4	Details of Experiments reported in Fig. 2.1f for CIFAR-10	115
B.1	Details of Experiments reported for MNIST with MLP	133
B.2	Details of Experiments reported for MNIST with CNN	133
B.3	Details of Experiments reported for Fashion-MNIST with CNN	133
B.4	Details of Experiments reported for CIFAR10 with CNN	135

List of Figures

1.1	Learning Setup	4
2.1	Numerical results for various datasets and architectures. All x -axes show the number of Epochs of training. Fig. 2.1a shows the effect of different amounts of heldout data on the summands appearing in our bound, and what those would be if we upper bounded the <i>incoherence</i> $\ \xi\ $ by $\ \nabla\hat{R}\ $ when it is not 0. Fig. 2.1b compares a Monte Carlo estimate of our bound with that of [MWZZ18] and shows the effect of inverse temperature on each. Fig. 2.1c compares a Monte Carlo estimate of our bound with that of [MWZZ18] and shows the effect of learning rate on each. Figs. 2.1d to 2.1f compare the summands appearing in our bound and those of [MWZZ18] across datasets.	21
3.1	Numerical results for various datasets and architectures. All the x -axes represent the training iteration. The plots in the first column depict a Monte Carlo estimate of our bounds with that of Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19] and J. Li, Luo, and Qiao [LLQ20]. The plots in the second column compare the mean of the <i>training set incoherence</i> in [NHDKR19] with the two-sample incoherence in our bound. Finally, the plots in the third column show the mean of the squared error probability of the hypothesis testing performed by the proposed prior in our bound.	37
5.1	A counter example.	61
5.2	Conditional independence relationships encoded as a graphical model.	67
B.1	Comparison between constants of Theorem 3.1.1, Theorem 3.1.3, and Theorem B.2.1 for the case $k \rightarrow \infty$	119
B.2	Learning curves. These plots show the training error, error on the test set, and the training loss. The loss functions is cross-entropy. Note y -axes for the error plots are log-scale.	134
C.1	One-inclusion graph of point functions for a set of distinct points.	147

D.1	The upper bound in Eq. (D.29)	172
D.2	Upper bound on $n - \text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ in Eq. (D.36).	176

Chapter 1

Introduction

Ensuring that Machine Learning (ML) models perform well during *test-time* is crucial for their reliable deployment in sensitive applications. In learning theory, this *requirement* can be formalized using the concept of *generalization error*. Generalization error captures the difference between a model’s performance on the available data (a.k.a training data) and its performance on unseen data. In this thesis, we investigate the problem of characterizing generalization error from an information-theoretic perspective. We aim to demonstrate how various notions of mutual information between the output of a learning algorithm and the training set can effectively capture the generalization error. Our approach holds the potential to provide novel insights into the trade-offs among various components of a learning problem, and guide the development of more robust and accurate machine learning models.

1.1 What is Learning from Data?

Assume have a known data space $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ where \mathcal{X} is the feature space and \mathcal{Y} is the label space. For instance, for MNIST data [LCB10], $\mathcal{X} = \mathbb{R}^{28 \times 28}$ and $\mathcal{Y} = \{0, \dots, 9\}$. Also, let $\ell : \mathcal{Z} \times \mathcal{Y}^{\mathcal{X}} \rightarrow \mathbb{R}$ denote a loss function where $\mathcal{Y}^{\mathcal{X}}$ is the set of all functions from \mathcal{X} to \mathcal{Y} . In particular for $h \in \mathcal{Y}^{\mathcal{X}}$ and $z = (x, y) \in \mathcal{Z}$, $\ell(z, h) = \ell((x, y), h)$ quantifies how *close* $h(x)$ is to y . Let \mathcal{D} be an *unknown* data distribution on the space \mathcal{Z} . For every $h \in \mathcal{Y}^{\mathcal{X}}$, let $R_{\mathcal{D}}(h) = \mathbb{E}_{Z \sim \mathcal{D}} [\ell(Z, h)]$ denote the population error of h . For a training set $S_n \in \mathcal{Z}^n$, the training error of h with respect to S_n is defined as $\hat{R}_{S_n}(h) = \frac{1}{n} \sum_{z \in S_n} \ell(h, z)$.

The problem of *learning from data* can be defined as follows. We assume a learner has access to an IID sample $S_n = (Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$ from the unknown data distribution \mathcal{D} . The learner is tasked to design a (potentially randomized) learning

algorithm $\mathcal{A} = (\mathcal{A}_n)_{n \geq 1}$, where \mathcal{A}_n is a mapping from \mathcal{Z}^n to $\mathcal{Y}^{\mathcal{X}}$, with the goal of minimizing $R_{\mathcal{D}}(\mathcal{A}_n(S_n))$. (See Fig. 1.1.). This quantity is known as population error of \mathcal{A}_n . A direct analysis of the population error is challenging. Also, it may give little understanding of the common properties among different learning problems which ensures low population error.

One approach to analyze the population error is to use the following decomposition

$$\text{Population Error of } \mathcal{A}_n = \text{Training Error of } \mathcal{A}_n + \text{Generalization Error of } \mathcal{A}_n,$$

where the difference between the population risk and the training set is defined as (expected) generalization error:

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) = \mathbb{E} \left[R_{\mathcal{D}}(\mathcal{A}_n(S_n)) - \hat{R}_{S_n}(\mathcal{A}_n(S_n)) \right]. \quad (1.1)$$

This fundamental, yet trivial, bound implies that a control on the generalization error and the training error let us bound the population error. Here, the training error corresponds to the empirical optimization gap, which can be bounded by standard optimization convergence analysis. The main focus of this thesis is on providing new techniques and tools to analyze $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$.

Remark 1.1.1. Indeed, in modern machine learning, it is possible for algorithms to achieve a training error of zero while still having a small test error. This phenomenon is known as *interpolation*. In such cases, the traditional decomposition of test error into training error and generalization error does not simplify the problem of estimating test error. In other words, the problem of charactering the generalization error is as difficult as the problem of charactering the test error. We will return to this point in Chapters 4 and 5. ◁

1.2 Information-Theoretic Measures for Generalization

In statistical learning theory, various approaches have been developed to analyze generalization error of learning algorithms. Many of these approaches rely on the concept of *stability*. Informally, a learning algorithm is said to be stable if a small change to its training set does not change the output of the algorithm much. An essential viewpoint for developing information-theoretic notions of stability is viewing ML algorithms as *stochastic* transformations that map training data to an output. In

this thesis, we are interested in understanding of

1. How can we quantify the stability of the distribution of the ML algorithms output? We refer to this type of stability as *distributional stability*.
2. What is the connection of distributional stability to the generalization error?

The field of information theory offers many beautiful mathematical tools that can help us measure stability. From an information-theoretic point of view, stability can be naturally measured by the mutual information between the output of a learning algorithm and its training set. This intuition has been formalized in the work by [RZ16; XR17; NHDKR19; BZV20b; HNKRD20; SZ20a], where the authors show that the expected generalization error can be bounded by

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) = O\left(\sqrt{\frac{\text{I}(\text{summary of } S_n; \text{summary of } \mathcal{A}_n(S_n))}{n}}\right),$$

where $\text{I}(\cdot; \cdot)$ denotes the mutual information. This result formalizes the intuition that a learning algorithm without heavy dependence on the training set will generalize well. The mutual information term in the upper bound is referred to as the *information complexity of \mathcal{A}_n* . However, it is worth noting that there is no unique way to define the information complexity, and previous research [RZ16; XR17; NHDKR19; BZV20b; HNKRD20; SZ20a] provides various definitions for it.

Remark 1.2.1. It is indeed worth mentioning as a historical note that mutual information-based generalization bounds were initially discovered in the PAC-Bayesian literature by researchers such as Catoni [Cat07] and Lever, Laviolette, and J. Shawe-Taylor [LLS13]. These early contributions explored the connection between mutual information and generalization error. However, despite their significance, these contributions were somewhat overlooked and remained unrecognized for a period of time. The full potential and implications of using mutual information as a measure for generalization were not widely appreciated until recently. It was the work of [RZ16] and [XR17] that revived and brought attention to these mutual information-based generalization bounds. ◁

Using information-theoretic measures to characterize generalization error in machine learning has notable advantages. One of the shortcomings of the several classical bounds is that they focus on the worst-case guarantees. For instance, the methods based on Vapnik–Chervonenkis theory [VC74], sample compression [LW86], and stability [BE02] provide generalization guarantee for the *worst-case* data distribution.

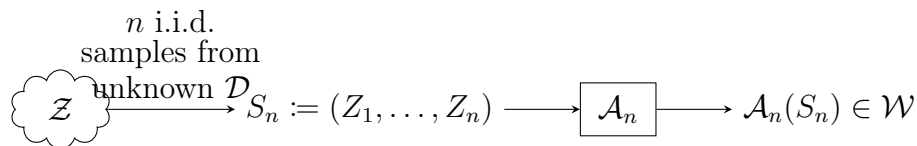


Figure 1.1: Learning Setup

This limitation severely restricts their ability to explain the success of modern ML algorithms, which often rely on natural data distributions. This issue is highlighted by [ZBHRV17] who conduct several interesting numerical experiments to study the effectiveness of the classical generalization bounds in explaining the success of deep neural networks (DNNs). Their key finding is that in the overparametrized regime, where the number of training points is much smaller the number of parameters of the DNN, the classical frameworks defined based on the worst-case complexity of parameter space are *unable* to explain the success of DNNs. In contrast, information-theoretic methods offer a significant advantage in that the generalization bound takes into account all the components of a learning problem, thereby enabling us to reason about generalization in an *instance-dependent* manner.

1.3 Main Questions and Overview of the Results

In this section, we present the motivation behind our study and an overview of the results in the thesis.

1.3.1 Characterizing Generalization Error of Iterative Learning Algorithms

As discussed in Section 1.2, information-theoretic methods demonstrate that the information complexity of learning algorithms plays a crucial role in controlling generalization error. This raises a natural question:

How can we characterize information complexity to investigate the generalization properties of practical learning algorithms such as iterative methods based on gradient descent, and what novel insights can we gain from this approach?

Despite the potential benefits of information-theoretic generalization bounds, there are several roadblocks to their application. One of the key roadblocks towards using information-theoretic frameworks for explaining generalization of iterative learning algorithm is that, in practice, we do not have access to the data distribution and the

information-theoretic measures depend on the data distribution. Instead, we only have a finite number of samples (a.k.a a training set) from the *unknown* data distribution. The next challenge is that it is not clear how to compute the mutual information terms underlying the information complexity measures since for many algorithms we do not have an analytical expression.

In Chapters 2 and 3 we investigate the question of obtaining generalization bounds for the models trained with Stochastic Gradient Langevin Dynamics (SGLD). SGLD is obtained by adding a isotropic Gaussian noise to the update rule of SGD. By introducing a small amount of Gaussian noise into the optimization process, SGLD allows for exploration of the parameter space while maintaining the ability to converge to a solution. In Chapter 2, our main contributions are significantly improved information complexity bounds for SGLD via data-dependent estimates. Our approach is based on the variational characterization of mutual information and the use of data-dependent priors that forecast the mini-batch gradient based on a subset of the training samples. This technique effectively casts the problem of characterizing the information complexity as a *two-player game*. Roughly speaking, we show that

$$\text{information complexity of SGLD} \leq \sum_{t=1}^T \text{Training Set Gradient Incoherence}_t,$$

where *Training Set Gradient Incoherence* at iteration t captures the variance of the gradient vectors of the training set at iteration t . We show that this quantity is typically orders of magnitude smaller than the squared gradient norms or Lipschitz constants that other bounds in the literature depend upon. Moreover, Our bound can be tied to a measure of flatness of the empirical risk surface.

One of the limitations of our results in Chapter 2 is that upperbound on generalization error in terms of the training set gradient incoherence may converge to infinity as $t \rightarrow \infty$, while the actual generalization gap remains constant. To address this issue, in Chapter 3 we study the recently proposed information complexity measure by Steinke and Zakyntinou [SZ20a], to reason about the generalization error of a learning algorithm by introducing a super sample that contains the training sample as a random subset and computing mutual information conditional on the super sample. We introduce tighter bounds compared to the bounds in [SZ20a], building on the *individual sample* idea of [BZV20b] and the *data dependent* ideas from Chapter 2, using disintegrated mutual information. Finally, we apply these bounds to the study of Langevin dynamics algorithm, showing that conditioning on the super sample allows

us to exploit information in the optimization trajectory to obtain tighter bounds based on hypothesis tests. Roughly speaking we show that

$$\text{information complexity of SGLD} \leq \sum_{t=1}^T \alpha_t \cdot \text{Two Sample Incoherence}_t,$$

where *two sample incoherence* at iteration t quantifies the discrepancy between the gradient of training set and the gradient on a fresh new sample. Also, $0 < \alpha_t < 1$ and usually it converges to zero as $t \rightarrow \infty$. The distinguishing feature of our data-dependent bound in Chapter 3 is that the penalty terms get “filtered” α_t which yields bounds that do not goes to infinity even when $t \rightarrow \infty$.

A significant recurring theme among Chapters 2 and 3 is that we upperbound the information complexity of SGLD using certain statistics of the optimization trajectory. Moreover, this technique reduces the problem of predicting the generalization error to the estimation of scalar random variables such as variance of the training examples’ gradients. Also, our technique does not place restrictions on the learning rate or Lipschitz continuity of the loss or its gradient and can provide a *distribution-dependent* generalization error.

A natural question that follows is whether we can use the techniques developed for SGLD in Chapters 2 and 3 to study the generalization error of Stochastic Gradient Descent (SGD). We provide an affirmative answer to this question in [NDHR21] by introducing the idea of *surrogate algorithm*. We construct an explicit surrogate algorithm that is coupled to SGD and closely *mimics* its dynamics, but for which the generalization performance is amenable to an information-theoretic analysis. Our generalization bounds for SGD offer new insights on the impact of flatness of the loss landscape as well as the variance of the training examples’ gradient on the performance of SGD. We provide a summary of this result in Chapter 6.

In summary, the results presented in Chapter 2 and Chapter 3 demonstrate that we can derive generalization error bounds for the SGLD algorithm by analyzing its information complexity, taking into account both the distribution and algorithmic properties. Furthermore, our findings reveal that specific aspects of the optimization trajectory have an impact on the generalization performance of SGLD. This observation suggests that these aspects could potentially serve as implicit biases of the algorithm, influencing its ability to generalize well beyond the training data.

1.3.2 Unifying Framework for Generalization

Information-theoretic generalization bounds are distribution- and algorithm-dependent, making them well-suited for studying generalization error of natural data distributions. In contrast, other generalization bounds, such as VC theory and uniform stability, can only be used to reason about worst-case generalization properties. This raises the following question

For which learning problems are information-theoretic bounds expressive enough to estimate the optimal minimax performance?

Affirmatively answering this question suggests that information complexity is an appropriate measure of complexity that can unify and connect different existing approaches. To address this fundamental question, we pose the following research problems: 1) For which learning problems and learning algorithms are information-theoretic bounds expressive enough to accurately estimate the optimal minimax generalization error? 2) What is the expressive power of information-theoretic generalization bounds?

In Chapters 4 to 6, we improve our understanding of the expressive power of information theoretic generalization frameworks in the context of binary classification for VC classes and gradient methods in the stochastic convex optimization.

We first investigate this problem in the context of binary classification with zero-one loss in Chapters 4 and 5. In Chapter 4, we investigate the expressiveness of the “conditional mutual information” (CMI) framework of Steinke and Zakynthinou [SZ20a] and the prospect of using it to provide a unified framework for proving generalization bounds in the realizable setting. We first demonstrate that one can use this framework to express non-trivial (but sub-optimal) bounds for any learning algorithm that outputs hypotheses from a class of bounded VC dimension. We then explore two directions of strengthening this bound: (i) Can the CMI framework express optimal bounds for VC classes? (ii) Can the CMI framework be used to analyze algorithms whose output hypothesis space is unrestricted (i.e. has an unbounded VC dimension)? With respect to Item (i) we prove that the CMI framework yields the optimal bound on the expected risk of *Support Vector Machines* (SVMs) for learning halfspaces. This result is an application of our general result showing that *stable* compression schemes [BHMZ20] of size k have uniformly bounded CMI of order $O(k)$. We further show that an inherent limitation of proper learning of VC classes contradicts the existence of a proper learner with constant CMI, and it implies a negative resolution to an open problem of Steinke and Zakynthinou [SZ20b]. We further study the CMI of empirical risk minimizers (ERMs) of class \mathcal{H} and show that it is possible to output

all consistent classifiers (version space) with bounded CMI *if and only if* \mathcal{H} has a bounded star number [HY15]. With respect to Item (ii) we prove a general reduction showing that “leave-one-out” analysis is expressible via the CMI framework. As a corollary we investigate the CMI of the one-inclusion-graph algorithm proposed by Haussler, Littlestone, and M. K. Warmuth [HLW94]. More generally, we show that the CMI framework is universal in the sense that for *every* consistent algorithm and data distribution, the expected risk vanishes as the number of samples diverges *if and only if* its evaluated CMI has sublinear growth with the number of samples.

Some of our results in Chapter 4 are suboptimal by a log factor. In Chapter 5, we propose a new information complexity measure called *leave-one-out conditional mutual information*: the mutual information between (certain summaries of) the output of a learning algorithm and its n training data, conditional on a supersample of $n + 1$ i.i.d. data from which the training data is chosen at random without replacement. These leave-one-out variants of the conditional mutual information (CMI) of an algorithm [SZ20a] are also seen to control the mean generalization error of learning algorithms with bounded loss functions. For learning algorithms achieving zero empirical risk under 0–1 loss (i.e., interpolating algorithms), we provide an explicit connection between leave-one-out CMI and the classical leave-one-out error estimate of the risk. Using this connection, we obtain upper and lower bounds on risk in terms of the (evaluated) leave-one-out CMI. When the limiting risk is constant or decays polynomially, the bounds converge to within a constant factor of two. As an application, we analyze the population risk of the one-inclusion graph algorithm, a general-purpose transductive learning algorithm for VC classes in the realizable setting. Using leave-one-out CMI, we match the optimal bound for learning VC classes in the realizable setting, answering an open challenge raised by Steinke and Zakynthinou [SZ20a]. Finally, in order to understand the role of leave-one-out CMI in studying generalization, we place leave-one-out CMI in a hierarchy of measures, with a novel unconditional mutual information at the root. For 0–1 loss and interpolating learning algorithms, this mutual information is observed to be precisely the risk.

Despite these successful developments, much less is known about the optimality or limitations of information-theoretic frameworks beyond the setting of binary classification and zero-one loss function valued loss. In Chapter 6, we consider the prospect of establishing minimax rates for gradient descent in the setting of stochastic convex optimization via several existing information complexity measures: input-output mutual information bounds, conditional mutual information bounds and variants, PAC-Bayes bounds, and recent conditional variants thereof. We prove that none of these bounds

are able to establish minimax rates. We then consider a common tactic employed in studying gradient methods, whereby the final iterate is corrupted by Gaussian noise, producing a noisy “surrogate” algorithm. We prove that minimax rates cannot be established via the analysis of such surrogates. Our results suggest that new ideas are required to analyze gradient descent using information-theoretic techniques.

In summary, our results in Chapters 4 to 6 improve our understanding of the expressive power of information-theoretic generalization bounds. These chapters provide a comprehensive analysis of the strengths and limitations of these bounds, yielding both positive and negative results.

Chapter 2

Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates

2.1 Introduction

Stochastic subgradient methods, especially stochastic gradient descent (SGD), are at the core of recent advances in deep-learning practice. Despite some progress, developing a precise understanding of generalization error for that class of algorithms remains wide open. Concurrently, there has been steady progress for noisy variants of SGD, such as stochastic gradient Langevin dynamics (SGLD) [WT11; GM91; RRT17] and its full-batch counterpart, Langevin dynamic [GM91]. The introduction of Gaussian noise to the iterates of SGD expands the set of theoretical frameworks that can be brought to bear on the study of generalization. In pioneering work, Raginsky, Rakhlin, and Telgarsky [RRT17] exploit the fact that SGLD approximates Langevin diffusion, a continuous time Markov process, in the small step size limit. One drawback of this and related analyses involving Markov processes is the reliance on mixing: in order to obtain a non-trivial upperbound on the excess error, the number of iterations needs to be scaled with the dimension of the parameter space. We hypothesize that SGLD is not mixing in practice, so results based upon mixing may not be representative of empirical performance.

In recent work, Pensia, Jog, and Loh [PJL18] perform a stepwise analysis of a family of noisy iterative algorithms that includes SGLD and Langevin dynamic. At the foundation of this work is the framework of Russo and J. Zou [RZ15] and Xu

and Raginsky [XR17], where mean generalization error is controlled in terms of the mutual information between the dataset and the learned parameters. However, because the data distribution is unknown, so is any mutual information involving the data. This presents a significant barrier to understanding generalization in terms of mutual information.

One of the key contributions of Pensia et al. is a bound on the mutual information between the data and the final weights, which they construct from a bound on the mutual information between the data and the entire trajectory of weights. By exploiting properties of mutual information, they express the latter as a sum of conditional mutual informations associated with each gradient step. While these conditional mutual informations are also unknown, Pensia et al. obtain a bound in terms of the Lipschitz constant for the objective function being optimized.

By passing to the full trajectory and exploiting Lipschitz continuity, Pensia et al. circumvent the statistical barrier posed by the unknown mutual information. Their analysis, however, introduces several sources of looseness. In particular, the use of Lipschitz constants, which lead to distribution-*independent* bounds, eradicates any hope that these bounds will be non-vacuous for modern models and datasets. Indeed, for deep neural networks, the Lipschitz constant for the empirical risk would be prohibitively large, or in some cases infinite, and would immediately render any bound that depends on them vacuous in regimes of interest. In order to fully exploit the decomposition proposed by [PJJ18], one needs distribution-*dependent* bounds on the incremental mutual information at each step.

The key contribution of the present section is the observation that variants of the mutual information between the learned parameters and a subset of the data can be estimated using the rest of the data. We refer to such estimates as *data-dependent* due to their intermediate dependence on part of the data. The use of data-dependent estimates leads to distribution-dependent bounds that naturally adapt to the model of interest and the data distribution. In particular, using data-dependent estimates, we arrive at bounds in terms of the *incoherence* of gradients in the dataset. Roughly speaking, the incoherence measures the amount by which batch gradients computed on subsets of the data disagree, as quantified by squared norm. Crucially, the incoherence is never larger than the squared-gradient-norm on average, and the incoherence is 0 for most iterations of SGLD with small batches.

In the process of developing tighter distribution-dependent bounds, we also observe that, in some circumstances, one may obtain tighter estimates by working with conditional or disintegrated information-theoretic quantities. In particular, doing

so provides more opportunities to exchange expectation and concave functions than are available with previous mutual information bounds. Using their own mutual information bound and the chain rule, [BZV20a] improve on the generalization error bound for SGLD from [PJJ18] by a factor of $\sqrt{\log n}$ where n is the sample size. The advantage of [BZV20a] that enables this improvement is that their bound is only penalized once per epoch at a randomly chosen step. This effectively changes the order of an expectation and square-root, improving the bound. Building upon [BZV20a; RZ15; XR17], we develop generalization bounds in terms of disintegrated information-theoretic quantities that extract expectations from concave functions as much as possible.

We start this section by establishing some essential definitions.

2.1.1 Preliminaries

For random variables X and Y , write $\mathbb{E}^Y X = \mathbb{E}[X|Y]$ and $\mathbb{P}^Y[X]$ for the conditional expectation and (regular) conditional distribution, respectively, of X given Y .¹ Besides the usual notions of KL divergence, mutual information, and conditional mutual information (see Appendix A.1 for formal definitions), we rely on the following less common notion:

Definition 2.1.1. Let X , Y , and Z be arbitrary random elements. Let \otimes form product measures. The *disintegrated mutual information between X and Y given Z* is

$$I^Z(X; Y) = \text{KL}(\mathbb{P}^Z[(X, Y)] \| \mathbb{P}^Z[X] \otimes \mathbb{P}^Z[Y]).$$

It follows immediately from definitions that $I(X, Y|Z) = \mathbb{E}I^Z(X, Y)$. Letting ϕ satisfy $\phi(Z) = I^Z(X; Y)$ a.s., define $I(X, Y|Z = z) = \phi(z)$. This notation is necessarily well defined only up to a null set under the marginal distribution of Z .

2.2 Methods

In this section, we establish generalization bounds for learning algorithms in terms of information-theoretic quantities that depend on the unknown data distribution and the probabilistic properties of the learning algorithm. We then describe two complementary strategies that we employ to bound these otherwise intractable quantities. In Section 6.4, we apply these methods to the study of Langevin dynamic and SGLD.

¹We fix arbitrary versions and assume regular versions of conditional distributions exist.

We make repeated use of generalized notions of *priors* and *posteriors*, which arise in the PAC-Bayes literature ([Cat07; SW97; McA99b], etc.) and relate to variational bounds on mutual information, which we will now describe: Consider learned parameters W , data S , and auxiliary variables V , viewed as random elements in \mathcal{W} , Z^n , etc., respectively. In PAC-Bayes, a generalized posterior is an arbitrary random measure on \mathcal{W} . In our setting, the *posterior*, Q , (of W given S and V) is the conditional distribution of W given S and V . (Formally, Q is a probability kernel, but one can think informally that $Q = f(S, V)$ for some measurable function taking values in the space of Borel probability measures, and so we will simply say that Q is $\sigma(S, V)$ -measurable.)

Definition 2.2.1 (Data-dependent prior). Let Q be a $\sigma(S, V)$ -measurable posterior. A (generalized) *prior* P is a random measure on \mathcal{W} , measurable with respect to some sub- σ -algebra of $\sigma(S, V)$. A prior P is said to be *data-dependent* if it is not independent of S .

Let P be a \mathcal{F} -measurable data-dependent prior, where $\sigma(V) \subset \mathcal{F}$. Using a variational characterization of mutual information (see Appendix A.2.1), we have

$$\mathbb{E}^{\mathcal{F}}[\text{KL}(Q \parallel P)] \geq I^{\mathcal{F}}(W; S) \text{ a.s.}, \quad (2.1)$$

with equality for $P = \mathbb{P}^{\mathcal{F}}[W]$. Therefore, if the expected KL divergence is small, W contains little information about S beyond what is already captured by \mathcal{F} . If the special case where the disintegrated mutual information is zero, then W is independent of S given \mathcal{F} . In the context of generalization, this implies that the data S not contained in \mathcal{F} can be used to form an unbiased estimate of the risk of W . The bounds we present below extend this logic to nonzero mutual information.

The utility of using data-dependent priors to control disintegrated mutual information depends on the balance of two effects: On the one hand, $I(W; S) \leq I(W; S | \mathcal{F})$ since \mathcal{F} independent of the training set, and so conditioning never improves a theoretical bound and may make it looser. On the other hand, $I(W; S)$ depends on the *unknown* data distribution and so distribution-independent bounds will often be very loose. In contrast, the KL divergence based on P can exploit the information in $\mathcal{F} \subset \sigma(S, V)$ to obtain tighter data-dependent bounds on $I^{\mathcal{F}}(W; S)$.

In order to construct data-dependent priors, we partition the dataset S in two halves, based on a random subset $J \subset \{1, \dots, n\}$ with $|J| = m$ nonrandom. Let $J = \{j_1, \dots, j_m\}$, The first half, $S_J = (Z_{j_1}, \dots, Z_{j_m})$, contains m points, which we will use to construct a data-dependent prior P . The second half, S_J^c , containing the

remaining $n - m$ points, is independent of P . (Note that S_J and S_J^c are independent of J , since m is nonrandom.)

This particular construction of data-dependent priors allow us to leverage a type of *non-uniform KL-stability*: the prior P may exploit S_J to make a data-dependent forecast of Q , yielding a bound, B , on the conditional expected generalization error (with respect to the remaining $n - m$ data points in S_J^c). Averaging over S_J , we obtain a bound on the (unconditional) expected generalization error.

Definition 2.2.2. Let S_J, S_J^c be defined as above. Suppose that \mathcal{F} is a σ -field with $\sigma(S_J) \subset \mathcal{F} \perp \sigma(S_J^c)$. An expected generalization error bound based on a *data-dependent estimate* is one of the form

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \mathbb{E}[B], \quad (2.2)$$

where B is \mathcal{F} measurable, and satisfies $\mathbb{E}^{\mathcal{F}}[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W)] \leq B$.

The idea of using data-dependent priors to obtain tighter bounds is standard in the PAC-Bayes literature [APS07; PASS12; DR18a; RSSPS18], but its utility in the present work is brought through by our introduction of data-dependent estimates. In the following section, we derive information-theoretic bounds on expected generalization error that can exploit data-dependent priors to form data-dependent estimates. We will then use these tools to study SGLD, without mixing assumptions.

2.2.1 Information-Theoretic Generalization Bounds based on Random Subsets of Data

Existing work by Xu and Raginsky [XR17] bounds the expected generalization error of a learning algorithm in terms of the mutual information between the random parameters and the data. The following result is a simple extension of [XR17, Thm. 1] that bounds the expected generalization error in terms of the mutual information between the parameters and a random subset of the data.

Theorem 2.2.3 (Data-Dependent Mutual Information Bound). *Let W be a random element in \mathcal{W} , let $S \sim \mathcal{D}^n$, and let $J \subseteq [n]$, $|J| = m$, be uniformly distributed and independent from S and W . Suppose that $\ell(Z, w)$ is σ -subgaussian when $Z \sim \mathcal{D}$, for each $w \in \mathcal{W}$. Let $Q = \mathbb{P}^S[W]$, and let P be a $\sigma(S_J)$ -measurable data-dependent prior*

on \mathcal{W} . Then

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \sqrt{2 \frac{\sigma^2}{n-m} I(W; S_J^c)} \leq \sqrt{2 \frac{\sigma^2}{n-m} \mathbb{E}[\text{KL}(Q \parallel P)]}.$$

The proof of this result can be found in Appendix A.2. When $m = 0$, this recovers [XR17, Thm. 1].

When the size of the subset is $m = n - 1$, this bound is weaker than [BZV20a, Prop. 1], due to the order of the concave square-root function and the expectation over the choice datapoint to be left out. This difference is addressed by our next result.

Randomization is one way that learning algorithms can control the mutual information between (a random subsets of) the data and the learned parameter. Let U be a random element independent from S and J , representing some aspect of the source of randomness used by the learning algorithm. Because $S \perp\!\!\!\perp \{J, U\}$ and $S \sim \mathcal{D}^n$, we have $(S_J, U) \perp\!\!\!\perp S_J^c$ and thus

$$I(W; S_J^c) \leq I(W; S_J^c | S_J, U) = \mathbb{E} I^{S_J, U}(W; S_J^c),$$

where the last equality follows from the definition of conditional mutual information. The next result shows that we can pull the expectation over both S_J and U outside the concave square-root function. In the case of SGLD, U will be the sequence of minibatch index sets.

Theorem 2.2.4 (Data-Dependent Disintegrated Mutual Information Bound). *Let W , S , and J be as in Theorem 2.2.3, and let U be independent from S and J . Suppose that $\ell(Z, w)$ is σ -subgaussian when $Z \sim \mathcal{D}$, for each $w \in \mathcal{W}$. Let $Q = \mathbb{P}^{S, U}[W]$ and let P be a $\sigma(S_J, U)$ -measurable data-dependent prior on \mathcal{W} . Then*

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \mathbb{E} \sqrt{2 \frac{\sigma^2}{n-m} I^{S_J, U}(W; S_J^c)} \leq \mathbb{E} \sqrt{2 \frac{\sigma^2}{n-m} \mathbb{E}^{S_J, U} \text{KL}(Q \parallel P)}$$

The proof of this result can be found in Appendix A.2. Since $I^{S_J, U}(W; S_J^c)$ is (S_J, U) -measurable, we may use S_J and U to obtain a data-dependent bound. In the case that $m = n - 1$, our bound is similar to, but not strictly comparable to, [BZV20a, Prop. 1]. Our bound is incomparable due to our use of disintegrated mutual information, $I^{S_J}(W; S_J^c)$ and the fact that we take the expectations over the dataset outside of the convex square-root function. The disintegrated mutual information cannot be upper bounded by the full mutual information, $I(W, S_J^c)$, which appears in [BZV20a] (even by taking expectations under the square root using Jensen's inequality).

However, Theorem 2.2.4 is essentially a disintegrated version of [BZV20a, Prop. 1]. In their actual SGLD expected generalization error bound, [BZV20a] controls the unconditional mutual information using the Lipschitz constant of the surrogate loss. Hence, one could easily recover the same bound using our result. The conditioning we have done, however, allows us to control the mutual information more carefully in order to achieve a tighter bound for SGLD than is provided by [BZV20a].

These bounds allow for a tradeoff: for large m , the mutual information is measured between the parameter and a small random subset of the data, and so we expect the mutual information to be small. (Indeed, this term will decrease monotonically in m .) At the same time, the $\frac{1}{n-m}$ term is larger, reflecting the reduced effect of averaging over only $n - m$ data to form our estimate of the empirical risk. It is unclear without further context whether this bound is tighter in the regime of small, intermediate, and large m . In fact, we find that, for the bounds we derive in our applications, $m = n - 1$ is optimal. This difference materially affects the quality and tightness of the bounds, as is discussed in Remark 2.3.4. However, for $m = n - 1$ and bounded loss, the following bound is tighter, while it is incomparable for other values of m .

Theorem 2.2.5 (Data-Dependent KL Bound). *Let W , S , J , and U be as in Theorem 2.2.4. Let $Q = \mathbb{P}^{S,U}[W]$ and let P be a $\sigma(S_J, U)$ -measurable data-dependent prior on \mathcal{W} . Suppose that $\ell(Z, w)$ is $[a_1, a_2]$ -bounded a.s. when $Z \sim \mathcal{D}$, for each $w \in \mathcal{W}$.*

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \mathbb{E} \sqrt{\frac{(a_2 - a_1)^2}{2} \text{KL}(Q \| P)}.$$

The proof of this result can be found in Appendix A.2. For an analytic comparison of the three bounds in the case that $m = n - 1$, see Appendix A.6. Remark A.2.2 explains why this result is only stated for bounded loss functions.

2.2.2 Decomposing KL Divergences and Mutual Information for Sequential Algorithms

Consider an iterative learning algorithm, and let $W_0, W_1, W_2, \dots, W_T \in \mathcal{W}$ be the parameters during the course of T iterations. In light of the variational bound for mutual information, we can obtain a generalization bound for W_T by bounding the expected KL divergences between the conditional distribution $\mathbb{P}^{S_J}[W_T]$ and some S_J -measurable “prior” distribution $P(Z)$. Unfortunately, the first distribution has no known tractable representation. Pensia, Jog, and Loh [PJL18] use monotonicity to bound a mutual information involving the terminal parameter with one involving the

full trajectory, then use the chain rule to decompose this into a sum of conditional mutual informations. The same principles allow us to first bound the terminal KL divergence by the KL for the full trajectory, and then decompose the KL divergence for the full trajectory over each individual step.

Setting some notation, let T be a nonnegative integer, let $[T]_0 = \{0, 1, 2, \dots, T\}$, let μ be a distribution on $\mathcal{W}^{[T]_0}$, and let X be a random variable with distribution μ . We are interested in naming certain marginal and conditional distributions (disintegrations) related to μ . In particular, for $t \in [T]_0$, let

- i) $\mu_t = \mathbb{P}[X_t]$, the marginal law of X_t ;
- ii) $\mu_{t|} = \mathbb{P}^{X_{0:(t-1)}}[X_t]$, the conditional law of X_t given $X_{0:(t-1)}$; and
- iii) $\mu_{0:t} = \mathbb{P}[X_{0:t}]$, the marginal law of $X_{0:t}$.

Proposition 2.2.6 (Decomposition of KL Divergences). *Let Q, P be probability measures on $\mathcal{W}^{[T]_0}$. Suppose that $Q_0 = P_0$. Then*

$$\text{KL}(Q_T \parallel P_T) \leq \text{KL}(Q \parallel P) = \sum_{t=1}^T \mathbb{E}_{Q_{0:(t-1)}}[\text{KL}(Q_t \parallel P_t)].$$

where, as per Section 2.1.1, Q_t is the conditional law of t -th iterate given the previous iterates, and so $\text{KL}(Q_t \parallel P_t)$ is a random variable which depends the $(W_0, \dots, W_{t-1}) \sim Q_{0:t-1}$.

The proof of this result may be found in Appendix A.2.

Considering the KL between full trajectories may yield a loose upper bound on the KL between terminal parameters (in particular, when the trajectory cannot be inferred from the terminus). We gain, however, analytical tractability, as we will see in the next section when we analyze particular algorithms stepwise. In fact, many bounds that appear in the literature implicitly require this form of incrementation. Our approach based on the KL divergence and data-dependent priors gives us much tighter control of the KL divergence contribution of each step.

2.3 Generalization Bounds for Specific Algorithms

Now that we have all of the theoretical tools required, we may establish bounds on the generalization error of specific noisy iterative learning algorithms by inventing sensible data-dependent priors. The use of a data-dependent prior which closely forecasts the true algorithm in each step is key in establishing tighter generalization bounds. We first consider the stochastic gradient Langevin dynamics (SGLD) algorithm [WT11],

then handle its full batch counterpart (unadjusted) Langevin dynamic [Erm75; DM17], which we will refer to LD. Note that the loss and risk functions used for training, $(\tilde{\ell}, \tilde{R}_{\mathcal{D}}, \tilde{R}_S)$, need not be the same loss functions used for assessing performance and generalization error, $(\ell, R_{\mathcal{D}}, \hat{R}_S)$, as explained in Section 2.1.1.

2.3.1 Stochastic Gradient Langevin Dynamics

Let η_t to be the learning rate at time t ; β_t be the inverse temperature at time t ; and ϵ_t , i.i.d. $\mathcal{N}(0, \mathbb{I}_d)$. Let b_t be the minibatch size at time t . We are interested in stochastic gradient Langevin dynamics, whose iterates are given by

$$W_{t+1} = W_t - \eta_t \nabla \tilde{R}_{S_t}(W_t) + \sqrt{2\eta_t/\beta_t} \epsilon_t. \quad (2.3)$$

where $\tilde{R}_{S_t}(w) = \frac{1}{b_t} \sum_{z \in S_t} \tilde{\ell}(w, z)$, and S_t is a subset of S of size b_t sampled uniformly at random with a sampling procedure which is independent of S , and independent of $\{\epsilon_t\}_{t \geq 0}$. The b_t data points in S_t are chosen *without replacement*.

A data-dependent prior for SGLD

Let S_J be a random subset of S , of size m , chosen independently from W_0, W_1, \dots , and independently of the sequence of minibatches, $\{S_t\}_{t \geq 0}$. Let the set of indices appearing in the t -th minibatch be denoted by K_t , so that $S_t = S_{K_t}$ for each t . By assumption, each K_t is a uniformly random subset of $\{1, \dots, n\}$ of size b_t . We set $U = (K_1, \dots, K_T)$, as to match the notation in the theorems of Section 2.2.1. Let $S_{J_t} = S_J \cap S_t = S_{J \cap K_t}$ and let $b'_t = |S_{J_t}|$. Let $S_t^c = S_t \setminus S_J = S_{K_t \setminus J}$ and $b_t^c = b_t - b'_t$. Define

$$\xi_t = \frac{b_t^c}{b_t} \left(\nabla \tilde{R}_{S_t^c}(W_t) - \nabla \tilde{R}_{S_J}(W_t) \right). \quad (2.4)$$

Let $Q(S, U)$ be the joint law of (W_0, \dots, W_T) given a dataset S and minibatch sequence U . Then $Q(S, U)$ is a random measure as it depends on the random dataset S and the sequence of indices U . It follows from Eq. (2.3) that $Q(S, U)_{t|}$ is multivariate normal with mean $\mu_{Q,t}(S, U) = W_t - \eta_t \nabla \tilde{R}_S(W_t)$ and covariance $2\frac{\eta_t}{\beta_t} \mathbb{I}_d$. Consider the data-dependent prior defined so that its conditional $P_{t|}(S_J, U)$ is a multivariate normal with covariance $2\frac{\eta_t}{\beta} \mathbb{I}_d$, and with mean

$$\mu_{P,t}(S_J, U) = W_t - \eta_t \left(\frac{b'_t}{b_t} \nabla \tilde{R}_{S_{J_t}}(W_t) + \frac{b_t - b'_t}{b_t} \nabla \tilde{R}_{S_J}(W_t) \right).$$

Note that $\mu_{Q,t}(S, U) - \mu_{P,t}(S_J, U) = \eta_t \xi_t(S, \text{idx})$. Thus the one-step KL divergence satisfies

$$2\text{KL}(Q_{t+1}|(S, \text{idx}) \parallel P_{t+1}|(S_J, U)) = \frac{\beta_t \eta_t}{4} \|\xi_t\|_2^2$$

Applying Proposition 2.2.6, we have (almost surely over the choice of (S, J, U))

$$2\text{KL}(Q_T(S, U) \parallel P_T(S_J, U)) \leq \sum_{t=1}^T \mathbb{E}^{S, J, U} \text{KL}(Q_t(S, U) \parallel P_t(S_J, U)) = \sum_{t=1}^T \mathbb{E}^{S, J, U} \frac{\beta_t \eta_t}{4} \|\xi_t\|_2^2.$$

Note that ξ_t depends on the exact weight sequence, and hence is $\sigma(S, J, U, W_{t-1})$ -measurable, but not $\sigma(S, J, U)$ -measurable. Hence, $\mathbb{E}^{S, J, U} \frac{\beta_t \eta_t}{8} \|\xi_t\|_2^2$ is a $\sigma(S, J, U)$ -measurable for each t .

Expected Generalization Error Bounds for SGLD

Theorem 2.3.1 (Expected Generalization Error Bounds for SGLD). *Let $\{W_t\}_{t \in [T]}$ denote the iterates of SGLD. Let the batch size be constant, $b_t = b$. If $\ell(Z, w)$ is σ -subgaussian for each $w \in \mathcal{W}$, then*

$$\mathbb{E}(R_{\mathcal{D}}(W_T) - R_S(W_T)) \leq \mathbb{E} \sqrt{\frac{\sigma^2}{n-m} \sum_{t=1}^T \frac{\beta_t \eta_t}{4} \mathbb{E}^{S, J, U} \|\xi_t\|_2^2} \leq \frac{\sigma}{2} \sqrt{\frac{n}{(n-1)^2} \sum_{t=1}^T \left(\frac{1}{b} + \frac{1}{n} \frac{n-m-1}{m}\right) \beta_t \eta_t \text{tr}(\mathbb{E}[\hat{\Sigma}_t(S)])} \quad (2.5)$$

and if $\ell(Z, w)$ is $[a_1, a_2]$ -bounded, and if $m = n - 1$, then

$$\mathbb{E}(R_{\mathcal{D}}(W_T) - R_S(W_T)) \leq \mathbb{E} \sqrt{\frac{(a_2 - a_1)^2}{4} \sum_{t=1}^T \frac{\beta_t \eta_t}{4} \mathbb{E}^{S, J, U} \|\xi_t\|_2^2} \leq \left[\frac{(a_2 - a_1)^2 n}{4(n-1)^2 b} \right]^{1/2} \mathbb{E} \sqrt{\sum_{t=1}^T \frac{\beta_t \eta_t}{4} \text{tr}(\mathbb{E}^S[\hat{\Sigma}_t(S)])} \quad (2.6)$$

where $\hat{\Sigma}_t(S) = \text{Var}_{Z \sim \text{Unif}(S)}^{W_t, S}(\nabla \tilde{R}_Z(W_t))$ is the finite population variance matrix of surrogate gradients.

Proof. The results are the direct combinations of Theorem 2.2.4 and Propositions 2.2.6 and A.2.1; and Theorem 2.2.5 and Proposition 2.2.6, respectively, with our data-dependent prior. Jensen's inequality is used to move expectations under $\sqrt{\cdot}$. Lemma A.4.2 expresses the results in terms of $\hat{\Sigma}$. \square

Remark 2.3.2. Suppose that $\beta_t = \beta$, $b_t = b$, and $m = n - 1$. Under uniform moment conditions on $\mathbb{E}^{S, J, U} \|\xi_t\|_2^2$, our generalization error bounds in Eq. (2.5) is clearly

$O(\sqrt{(\beta/bn) \sum_{t \leq T} \eta_t})$. Since $\xi_t = 0$ whenever $K_t \subset J$, we find that our first bound in Eq. (2.5) is also $O((1/n) \sum_{t \leq T} \sqrt{\beta \eta_t})$. To see this, notice that for non-negative random variables C_t and $B_t \sim \text{Ber}(p)$,

$$\mathbb{E} \sqrt{\sum_{t=1}^T B_t C_t} \leq \mathbb{E} [\sum_{t=1}^T B_t \sqrt{C_t}] = p \sum_{t=1}^T \mathbb{E} [\sqrt{C_t} | B_t = 1].$$

When $m = n - 1$, taking $B_t = I_{\xi_t \neq 0}$, $p = b/n$, $C_t = \frac{\beta_t \eta_t}{8} \mathbb{E}^{S_J, J, U} \|\xi_t\|_2^2$ yields the stated rate. \triangleleft

2.3.2 Langevin Dynamics

Under the same notation as above, the iterates of the Langevin dynamics algorithm are given by

$$W_{t+1} = W_t - \eta_t \nabla \tilde{R}_S(W_t) + \sqrt{2\eta_t/\beta_t} \varepsilon_t. \quad (2.7)$$

Expected Generalization Error Bounds for LD

We can recover bounds generalization error bounds for LD as a special case of SGLD when the batch size is the dataset size, $b_t = n$ for all t . The data-dependent prior is the same as for SGLD.

Theorem 2.3.3 (Expected Generalization Error Bounds for Langevin Dynamics).

Let $\{W_t\}_{t \in [T]}$ denote the iterates of the Langevin dynamics algorithm. If $\ell(Z, w)$ is σ -subgaussian for each $w \in \mathcal{W}$, then

$$\mathbb{E}(R_{\mathcal{D}}(W_T) - R_S(W_T)) \leq \sqrt{\frac{\sigma^2}{(n-1)m} \sum_{t=1}^T \frac{\beta_t \eta_t}{4} \mathbb{E} \text{tr}(\hat{\Sigma}_t(S))}, \quad (2.8)$$

and if $\ell(Z, w)$ is $[a_1, a_2]$ -bounded and $m = n - 1$, then

$$\mathbb{E}(R_{\mathcal{D}}(W_T) - R_S(W_T)) \leq \mathbb{E} \sqrt{\frac{(a_2 - a_1)^2}{4} \sum_{t=1}^T \frac{\beta_t \eta_t}{4} \mathbb{E}^{S_J} \|\xi_t\|_2^2} \leq \frac{a_2 - a_1}{2(n-1)} \mathbb{E} \sqrt{\sum_{t=1}^T \frac{\beta_t \eta_t}{4} \mathbb{E}^S \text{tr}(\hat{\Sigma}_t(S))},$$

where $\hat{\Sigma}_t(S) = \text{Var}_{Z \sim \text{Unif}(S)}^{W_t, S}(\nabla \tilde{R}_Z(W_t))$ is the finite population variance matrix of surrogate gradients.

For asymptotic properties of this bound when $\tilde{\ell}$ is L -Lipschitz, as in [PJJ18], see Appendix A.5. For a simple analytic worked example of mean estimation using

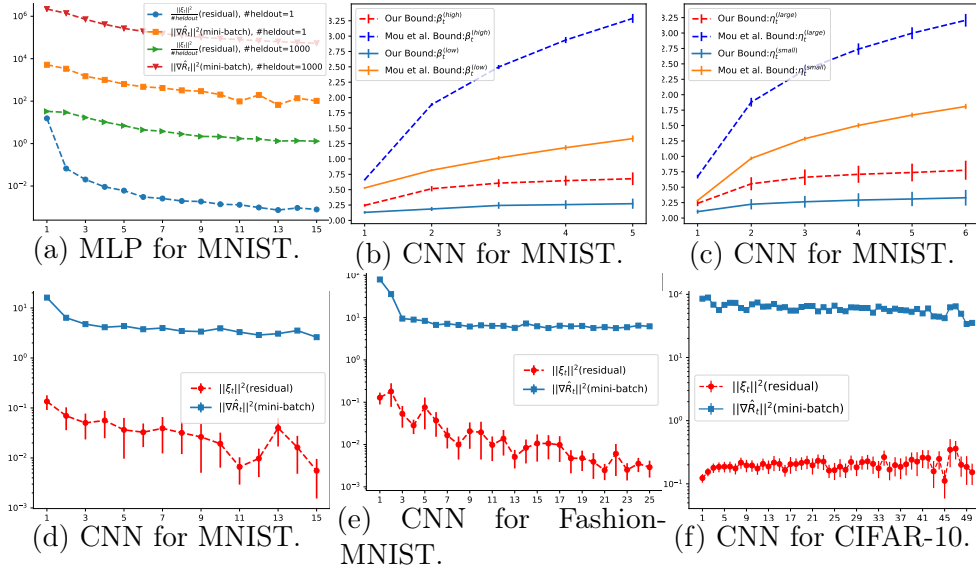


Figure 2.1: Numerical results for various datasets and architectures. All x -axes show the number of Epochs of training. Fig. 2.1a shows the effect of different amounts of heldout data on the summands appearing in our bound, and what those would be if we upper bounded the *incoherence* $\|\xi\|$ by $\|\nabla \tilde{R}\|$ when it is not 0. Fig. 2.1b compares a Monte Carlo estimate of our bound with that of [MWZZ18] and shows the effect of inverse temperature on each. Fig. 2.1c compares a Monte Carlo estimate of our bound with that of [MWZZ18] and shows the effect of learning rate on each. Figs. 2.1d to 2.1f compare the summands appearing in our bound and those of [MWZZ18] across datasets.

Langevin dynamics, refer to Appendix A.7.

Remark 2.3.4 (Dependence of our bounds on the subset size, m). The choice of $m \in \{1, \dots, n\}$ can make a material difference in the quality of the bound and whether it is vacuous or not. As seen in Eq. (2.8), if m is $\Omega(n)$ then the upper bound on expected generalization error is $O(\beta/n)$. If β is $\Omega(\sqrt{n})$, as is typical in practice, then overall, the bound is $O(n^{-1/2})$. If, on the other hand, m is $o(n)$ then the order of the bound with respect to n would be lower—in particular if m is $O(\sqrt{n})$ then our bound would not be decreasing in n for β of order $\Omega(\sqrt{n})$. \triangleleft

2.4 Empirical Results

We have developed bounds that depend on the gradient prediction residual of our data dependent priors (which we call the *incoherence* of the gradients), rather than on the gradient norms (as in [MWZZ18]) or Lipschitz constants (as in [PJL18; BZV20a]). The extent to which this represents an advance is, however, an empirical question. The functional form of our bounds and those in the cited work are nearly identical. The first key differences between our work and others is the replacement of gradient norms ($\|\nabla \tilde{R}_t\|^2$) and Lipschitz constants in other work with gradient prediction residual, ($\|\xi_t\|$) in our work. The second key difference is the order of expectations and square-

roots, which favor our bounds due to Jensen’s inequality. In this section, we perform an empirical comparison of the gradient prediction residual of our data dependent priors and the gradient norm across various architectures and datasets. This illustrates the first of the differences, the quantities appearing in the bound. Our results indicate that that our data-dependent priors yield significantly tighter results, as the sum of square gradient incoherences of our data dependent priors are between 10^2 and 10^4 times smaller than the sum of square gradient norms in the experiments we ran.

In Fig. 2.1, we compare $\|\xi_t\|^2$ and $\|\nabla\tilde{R}_t\|^2$ in order to assess the improvement our methods bring over existing results for SGLD. Specifically, the values of each plot are the averages of $\sqrt{\eta\beta}\|\xi_t\|/b$ and $\sqrt{\eta\beta}\|\nabla\tilde{R}_{S_t}\|/b$ over an epoch. These serve as estimates of the per-epoch contributions to the respective summations in our Theorem 2.3.1 and the bound of Mou, L. Wang, Zhai, and Zheng (Thm. 2 therein, when there is no L_2 -regularization). The average and standard error of both expressions taken over multiple runs are displayed. Bounds from related work that depend on Lipschitz constants would further upper bound what we show for [MWZZ18], by replacing $\|\nabla\tilde{R}_t\|$ with a Lipschitz constant. The Lipschitz constant could be lower bounded by the largest observed gradient norm, and would be off the chart.

From Fig. 2.1a, we see that the empirical performance reflects our analytical results that the bound is tighter for large m . As can be inferred from Eq. (2.4), the difference between $\|\xi_t\|^2$ and $\|\nabla\tilde{R}_t\|^2$ increases with m . From Figs. 2.1d to 2.1f we see that the squared gradient incoherence, $\|\xi_t\|^2$, are between 100 and 10,000 times smaller than the squared gradient norms, $\|\nabla\tilde{R}_t\|^2$ in all of these examples.

Using Monte Carlo simulation, we compared estimates of our expected generalization error bounds with (coupled) estimates of the bound from [MWZZ18]. The results, in Figs. 2.1b and 2.1c, show that our bounds are materially tighter, and remain non-vacuous after many more epochs. Fig. 2.1b also compares the two generalization error bounds for different inverse temperature schedules. Fig. 2.1c compares the two generalization error bounds based for different learning rate schedules. It can be inferred from Figs. 2.1b and 2.1c that our proposed bound yields to tighter values when the learning rate and the inverse temperature are small. However, it should be noted that with small learning rate and the inverse temperature, it would be difficult to have a very low training error when the empirical risk minimization is performed using SGLD.

The details of our model architectures, temperature, learning rate schedules and hyperparameter selections may be found in Appendix A.8. We did not aim to achieve the state-of-the art predictive performance. With further tuning, the prediction results

could be improved.

Chapter 3

Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms

3.1 Introduction

In this chapter, we study bounds on generalization error in terms of new class of information-theoretic bounds on generalization error, proposed by Steinke and Zakynthinou [SZ20a]. This approach was initiated by Russo and J. Zou [RZ15; RZ16]. The following result is due to Russo and J. Zou [RZ16] and Xu and Raginsky [XR17].

Theorem 3.1.1. $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \sqrt{\frac{\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)}{2n}}$.

Our focus in this chapter is on a new class of information-theoretic bounds on generalization error, proposed by Steinke and Zakynthinou [SZ20a]. Fix $k \geq 2$, let $[k] = \{1, \dots, k\}$, let $U^{(k)} = (U_1, \dots, U_n) \sim \text{Unif}([k]^n)$, and let $Z^{(k)} \sim \mathcal{D}^{\otimes(k \times n)}$ be a $k \times n$ array of IID random elements in \mathcal{Z} , independent from $U^{(k)}$. Let $S = (Z_{U_1,1}, \dots, Z_{U_n,n})$ and let W be a random element in \mathcal{W} such that conditional on S , $U^{(k)}$, and $Z^{(k)}$, W has distribution $\mathcal{A}_n(S)$. It follows that, conditional on S , W is independent from $U^{(k)}$ and $Z^{(k)}$. By construction, the data set S is hidden inside the super sample; the indices $U^{(k)}$ specify where. Steinke and Zakynthinou [SZ20a] use these additional structures to define:

Definition 3.1.2. The *conditional mutual information* of \mathcal{A}_n w.r.t. \mathcal{D} is

$$\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) = I(W; U^{(k)} | Z^{(k)}).$$

Intuitively, $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ captures how well we can recognize which samples from the given super-sample $Z^{(k)}$ were in the training set, given the learned parameters. This intuition and the connection of $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ with the membership attack [SSSS17] can be formalized using Fano’s inequality, showing that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ can be used to lower bound the error of any estimator of $U^{(k)}$ given W and $Z^{(k)}$. (See Appendix B.1.) Steinke and Zakynthinou [SZ20a] connect $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ with well-known notions in learning theory such as distributional stability, differential privacy, and VC dimension, and establish the following bound [SZ20a, Thm. 5.1] in the case $k = 2$, the extension to $k \geq 2$ being straightforward:

Theorem 3.1.3. $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \sqrt{\frac{2\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}}.$

This chapter improves our understanding of the framework introduced by Steinke and Zakynthinou [SZ20a], identifies tighter bounds, and applies these techniques to the analysis of a real algorithm. In Section 3.2, we present several formal connections between the two aforementioned information-theoretic approaches for studying generalization. Our first result bridges $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$, showing that for any learning algorithm, any data distribution, and any k , $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ is less than $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$. We also show that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ converges to $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ as $k \rightarrow \infty$ when $|\mathcal{W}|$ is finite. In Section 6.4, we establish two novel bounds on generalization error using the random index and super sample structure of Steinke and Zakynthinou, and show that both our bounds are tighter than those based on $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$. Finally, in Section 3.4, we show how to construct generalization error bounds for noisy, iterative algorithms using the generalization bound proposed in Section 6.4. Using the Langevin dynamics algorithm as our example, we introduce a new type of prior for iterative algorithms that “learns” from the past trajectory, using a form of *hypothesis testing*, in order to not “pay” again for information obtained at previous iterations. Experiments show that our new bound is tighter than [LLQ20; NHDKR19], especially in the late stages of training, where the hypothesis test component of the bound *discounts* the contributions of new gradients. Our new bounds are non-vacuous for a great deal more epochs than related work, and do not diverge or exceed 1 even when severe overfitting occurs.

3.1.1 Contributions

1. We characterize the connections between the $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$. We show that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ is always less than the $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ for any data distribution, learning algorithms and k . Further, we prove that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ converges to $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ when k goes to infinity for finite parameter spaces.
2. We provide novel generalization bounds that relate generalization to the mutual information between learned parameters and a random subset of the random indices U_1, \dots, U_n .
3. We apply our generalization bounds to the Langevin dynamics algorithm by constructing a specific *generalized prior and posterior*. We employ a generalized prior that learns about the values of the indices U from the optimization trajectory. To our knowledge, this is the first generalized prior that learns about the dataset from the iterates of the learning algorithm.
4. We show empirically that our bound on the expected generalization error of Langevin dynamics algorithm is tighter than other existing bounds in the literature.

3.1.2 Definitions from Probability and Information Theory

Let \mathcal{S}, \mathcal{T} be measurable spaces, let $\mathcal{M}_1(\mathcal{S})$ be the space of probability measures on \mathcal{S} , and define a probability kernel from \mathcal{S} to \mathcal{T} to be a measurable map from \mathcal{S} to $\mathcal{M}_1(\mathcal{T})$. For random elements X in \mathcal{S} and Y in \mathcal{T} , write $\mathbb{P}[X] \in \mathcal{M}_1(\mathcal{S})$ for the distribution of X and write $\mathbb{P}^Y[X]$ for (a regular version of) the conditional distribution of X given Y , viewed as a $\sigma(Y)$ -measurable random element in $\mathcal{M}_1(\mathcal{S})$. Recall that $\mathbb{P}^Y[X]$ is a regular version if, for some probability kernel κ from \mathcal{T} to \mathcal{S} , we have $\mathbb{P}^Y[X] = \kappa(Y)$ a.s. . If Y is $\sigma(X)$ -measurable then Y is a function of X . If random measure, P , is $\sigma(X)$ -measurable then the measure P is determined by X , but a random element Y with $\mathbb{P}^X[Y] = P$ is not X measurable unless it is degenerate. If X is a random variable, write $\mathbb{E}X$ for the expectation of X and write $\mathbb{E}^Y X$ or $\mathbb{E}[X|Y]$ for (an arbitrary version of) the conditional expectation of X given Y , which is Y -measurable. For a random element X on \mathcal{S} and a probability kernel P from \mathcal{S} to \mathcal{T} , the composition $P(X) := P \circ X$ is a $\sigma(X)$ -measurable random measure of a random element taking values in \mathcal{T} . We occasionally use this notation to refer to a kernel P implicitly by the way it acts on X .

Let P, Q be probability measures on a measurable space \mathcal{S} . For a P -integrable or nonnegative measurable function f , let $P[f] = \int f dP$. When Q is absolutely

continuous with respect to P , denoted $Q \ll P$, we write $\frac{dQ}{dP}$ for the Radon–Nikodym derivative of Q with respect to P . We rely on several notions from information theory: The *KL divergence of Q with respect to P* , denoted $\text{KL}(Q \parallel P)$, is $Q[\log \frac{dQ}{dP}]$ when $Q \ll P$ and ∞ otherwise. Let X, Y , and Z be random elements, and let \otimes form product measures. The *mutual information between X and Y* is $I(X; Y) = \text{KL}(\mathbb{P}[(X, Y)] \parallel \mathbb{P}[X] \otimes \mathbb{P}[Y])$. The *disintegrated mutual information between X and Y given Z* , is¹

$$I^Z(X; Y) = \text{KL}(\mathbb{P}^Z[(X, Y)] \parallel \mathbb{P}^Z[X] \otimes \mathbb{P}^Z[Y]).$$

The *conditional mutual information* of X and Y given Z is $I(X; Y|Z) = \mathbb{E}I^Z(X, Y)$.

3.2 Connections between $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$

In this section, we compare approaches for the information-theoretic analysis of generalization error, and we aim to unify the two main information-theoretic approaches for studying generalization. In Theorems 3.2.1 and 3.2.2 we will show that for any learning algorithm and any data distribution, $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ provides a tighter measure of dependence than $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$, and that one can recover $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ -based bounds from $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ for finite parameter spaces.

A fundamental difference between $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ is that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ is bounded by $n \log k$ [SZ20a], while $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ can be infinite even for learning algorithms that provably generalize [BZV20a]. One of the motivations of Steinke and Zakynthinou was that proper empirical risk minimization algorithms over threshold functions on \mathbb{R} have large $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ [BMNSY18]. In contrast, some such algorithms have small $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$. Our first result shows that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ is never larger than $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$.

Theorem 3.2.1. *For every $k \geq 2$, $I(W; S) = I(W; \tilde{Z}^{(k)}) + I(W; U^{(k)}|Z^{(k)})$ and*

$$\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \text{IOMI}_{\mathcal{D}}(\mathcal{A}_n).$$

Next, we address the role of the size of the super-sample in CMI. In [SZ20a], CMI is defined using a super-sample of size $2n$ ($k = 2$) only. Our next result demonstrates that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ agree $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ in the limit as $k \rightarrow \infty$ when the parameter space

¹ Letting ϕ satisfy $\phi(Z) = I^Z(X; Y)$ a.s., define $I(X, Y|Z = z) = \phi(z)$. This notation is necessarily well defined only up to a null set under the marginal distribution of Z .

is finite.

Theorem 3.2.2. *If the output of \mathcal{A}_n takes value in a finite set then*

$$\lim_{k \rightarrow \infty} \text{CMI}_{\mathcal{D}}(\mathcal{A}_n) = \text{IOMI}_{\mathcal{D}}(\mathcal{A}_n).$$

Combining Theorems 3.1.3 and 3.2.2, we obtain

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \lim_{k \rightarrow \infty} \sqrt{\frac{2\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}} = \sqrt{\frac{2\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}}, \quad (3.1)$$

when the parameter space is finite. Comparing Eq. (3.1) with Theorem 3.1.1 we observe that Eq. (3.1) is twice as large. In Theorem B.2.1, we present a refined bound based on $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ which asymptotically match Theorem 3.1.1. The proofs of the results of this section appear in Appendix B.3.

3.3 Sharpened Bounds based on Individual Samples

We now present two novel generalization bounds and show they provide a tighter characterization of the generalization error than Theorem 3.1.3. The results are inspired by the improvements on $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ due to Bu, S. Zou, and Veeravalli [BZV20a]. In particular, Theorem 3.3.1 bounds the expected generalization error in terms of the mutual information between the output parameter and a random subsequence of the indices $U^{(2)}$, given the super-sample. Theorem 3.3.4 provides a generalization bound in terms of the disintegrated mutual information between each individual element of $U^{(2)}$ and the output of the learning algorithm, W . The bound in Theorem 3.3.4 is an analogue of [BZV20a, Prop. 1] for Theorem 3.1.3. In this section as in Steinke and Zakyntinou [SZ20a], we only consider $Z^{(k)}$ and $U^{(k)}$ with $k = 2$, so we will drop the superscript from $U^{(k)}$. Let $U = (U_1, \dots, U_n)$. The proofs for the results of this section appear in Appendix B.4.

Theorem 3.3.1. *Fix $m \in [n]$ and let $J = (J_1, \dots, J_m)$ be a random subset of $[n]$, distributed uniformly among all subsets of size m and independent from W , $\tilde{Z}^{(2)}$, and U . Then*

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathbb{E} \sqrt{\frac{2I^{Z^{(2)}}(W; U_J | J)}{m}}. \quad (3.2)$$

By applying Jensen's inequality to Theorem 3.3.1, we obtain

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \sqrt{\frac{2I(W; U_J | Z^{(2)}, J)}{m}}. \quad (3.3)$$

Our next results in Theorem 3.3.2 let us compare Eq. (3.3) for different values of $m = |J|$.

Theorem 3.3.2. *Let $m_1 < m_2 \in [n]$, and let $J^{(m_1)}, J^{(m_2)}$ be random subsets of $[n]$, distributed uniformly among all subsets of size m_1 and m_2 , respectively, and independent from W , $\tilde{Z}^{(2)}$, and U . Then*

$$\frac{I(W; U_{J^{(m_1)}} | Z^{(2)}, J^{(m_1)})}{m_1} \leq \frac{I(W; U_{J^{(m_2)}} | Z^{(2)}, J^{(m_2)})}{m_2}. \quad (3.4)$$

Consequently, taking $m_2 = n$, for all $1 \leq m_1 \leq n$

$$\mathbb{E} \sqrt{\frac{2 I^{Z^{(2)}}(W; U_{J^{(m_1)}} | J^{(m_1)})}{m_1}} \leq \sqrt{\frac{2I(W; U | Z^{(2)})}{n}}. \quad (3.5)$$

Corollary 3.3.3. $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \sqrt{2 I(W; U_J | Z^{(2)}, J)/m}$. The case $m = |J| = n$ is equivalent to Theorem 3.1.3. The bound is increasing in $m \in [n]$, and, the tightest bound is achieved when $m = |J| = 1$. Also, Eq. (3.5) shows our bound in Theorem 3.3.1 is tighter than Theorem 3.1.3 for $k = 2$.

To further tighten Theorem 3.3.2 when $m = 1$, we show that we can pull the expectation over both $Z^{(2)}$ and J outside the concave square-root function.

Theorem 3.3.4. *Let $J \sim \text{Unif}([n])$ (i.e., $m = 1$ above) be independent from W , $\tilde{Z}^{(2)}$, and U . Then*

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathbb{E} \sqrt{2I^{Z^{(2)}, J}(W; U_J)} = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \sqrt{2I^{Z^{(2)}}(W; U_i)}. \quad (3.6)$$

Remark 3.3.5. Theorem 3.3.4 is tighter than Theorem 3.1.3 since

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \sqrt{2I^{Z^{(2)}}(W; U_i)} \leq \sqrt{\sum_{i=1}^n \frac{2}{n} I(W; U_i | Z^{(2)})} \leq \sqrt{\frac{2}{n} I(W; U | Z^{(2)})} \quad (3.7)$$

The first inequality is Jensen's, while the second follows from the independence of indices U_i . ◁

3.3.1 Controlling CMI bounds using KL Divergence

It is often difficult to compute MI directly. One standard approach in the literature is to bound MI by the expectation of the KL divergence of the conditional distribution of the parameters given the data (the “posterior”) with respect to a “prior”. The statement below is adapted from Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19].

Lemma 3.3.6. *Let X , Y , and Z be random elements. For all $\sigma(Z)$ -measurable random probability measures P on the space of Y ,*

$$I^Z(X; Y) \leq \mathbb{E}^Z[\text{KL}(\mathbb{P}^{X,Z}[Y] \parallel P)] \text{ a.s.}, \text{ with a.s. equality for } P = \mathbb{E}^Z[\mathbb{P}^{X,Z}[Y]] = \mathbb{P}^Z[Y].$$

We refer to the conditional law of W given S as the “posterior” of W given S , which we denote $Q = \mathbb{P}^S[W] = \mathbb{P}^{Z^{(2)},U}[W]$, and to P as the *prior*. This can be used in combination with, for example, Lemma 3.3.6 and Theorem 3.1.3 to obtain that for any $Z^{(2)}$ -measurable random prior $P(Z^{(2)})$

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \sqrt{\frac{2 I(W; U|Z^{(2)})}{n}} \leq \sqrt{\frac{2\mathbb{E}[\text{KL}(Q \parallel P(Z^{(2)}))]}{n}}. \quad (3.8)$$

Note that the prior only has access to $Z^{(2)}$, therefore from its perspective the training set can take 2^n different values. Alternatively, combining Lemma 3.3.6 and Theorem 3.3.1 yields

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathbb{E} \sqrt{\frac{2 \mathbb{E}^{Z^{(2)}} I^{Z^{(2)}}(W; U_J | U_{J^c}, J)}{m}} \leq \mathbb{E} \sqrt{\frac{2\mathbb{E}^{Z^{(2)}}[\text{KL}(Q \parallel P(Z^{(2)}, U_{J^c}, J))]}{m}}. \quad (3.9)$$

In Eq. (3.9) the prior has access to $n - m$ samples in the training set, S_{J^c} , because $Z_{U_{J^c}}^{(2)} = S_{J^c}$. However, since $Z^{(2)}$ is known to the prior, the training set can take only 2^m distinct values from the point of view of the prior in Eq. (3.9). This is a significant reduction in the amount of information that can be carried by the indexes in U_J about the output hypothesis. Consequently, priors can be designed to better exploit the dependence of the output hypothesis and the index set.

3.3.2 Tighter Generalization bound for the case $m = 1$

Since the strategy above controls MI-based expressions via KL divergences, one may ask whether a bound derived with similar tools, but directly in terms of KL, can be tighter than the combination Lemma 3.3.6 and Theorem 3.3.1. The following result

shows that for $m = 1$ a tighter bound can be derived by pulling the expectation over both U_{J^c} and J outside the concave square-root function.

Theorem 3.3.7. *Let $J \sim \text{Unif}([n])$ be independent from W , U , and $Z^{(2)}$. Let $Q = \mathbb{P}^{Z^{(2)}, U}[W]$ and P be a $\sigma(Z^{(2)}, U_{J^c}, J)$ -measurable random probability measure. Then*

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathbb{E} \sqrt{2 \text{KL}(Q \| P)}. \quad (3.10)$$

Here, the KL divergence is between two $\sigma(Z^{(2)}, J, U)$ -measurable random measures, so is random.

3.4 Generalization bounds for noisy, iterative algorithms

We apply this new class of generalization bounds to non-convex learning. We analyze the Langevin dynamics (LD) algorithm [GM91], following the analysis pioneered by Pensia, Jog, and Loh [PJL18]. The example we set here is a blueprint for building bounds for other iterative algorithms. Our approach is similar to the recent advances by Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19] and J. Li, Luo, and Qiao [LLQ20], employing data-dependent estimates to obtain easily simulated bounds. We find our new results allow us to exploit past iterates to obtain tighter bounds. The influence of past iterates is seen to take the form of a hypothesis test.

3.4.1 Bounding Generalization Error via Hypothesis Testing

The chain rule for KL divergence is a key ingredient of information-theoretic generalization error bounds for iterative algorithms [PJL18; NHDKR19; LLQ20; BZV20a]. $\mathcal{W}^{\{0, \dots, T\}}$ denotes the space of parameters generated by an iterative algorithm in T iterations. For any measure, ν , on $\mathcal{W}^{\{0, \dots, T\}}$, and $W \sim \nu$, let ν_0 denote the marginal law of W_0 , and ν_t denote the conditional law of W_t given $W_0 \dots W_{t-1}$.

Lemma 3.4.1 (Chain Rule for KL). *Let Q, P be probability measures on $\mathcal{W}^{\{0, \dots, T\}}$ with $Q_0 = P_0$. The following lemma bounds the KL divergence involving the posterior for the terminal parameter with one involving the sum of the KL divergences over each individual step of the trajectory. Then*

$$\text{KL}(Q_T \| P_T) \leq \text{KL}(Q \| P) = \sum_{t=1}^T Q_{0:(t-1)}[\text{KL}(Q_t \| P_t)]$$

The benefit of using the chain rule to analyze the iterative algorithm are two-fold: first, we gain analytical tractability; many bounds that appear in the literature implicitly require this form of incrementation [LLQ20; PJJ18; NHDKR19; BZV20a]. Second, and novel to the present work, the *information in the optimization trajectory can be exploited* to identify U from the history of W .

In order to understand how the prior may take advantage of information from the optimization trajectory, consider applying Lemma 3.4.1 to the KL term in Eq. (3.9). We have

$$\text{KL}(Q_T \parallel P_T(Z^{(2)}, U_{J^c}, J)) \leq \sum_{t=1}^T \mathbb{E}^{Z^{(2)}, U_{J^c}, J} [\text{KL}(Q_t \parallel P_t(Z^{(2)}, U_{J^c}, J))].$$

Here $P_t(Z^{(2)}, U_{J^c}, J)$ is a $\sigma(Z^{(2)}, U_{J^c}, J, W_{0:t-1})$ -measurable random probability measure. The prior may use U_{J^c} , $Z^{(2)}$, and J to reduce the number of possible values that U can take to $2^{|J|}$. Moreover, since U_J is constant during optimization, $W_0, W_1, W_2, \dots, W_{t-1}$ may leak some information about U_J , and the prior can use this information to tighten the bound by choosing a P_t that achieves small $\text{KL}(Q_t \parallel P_t)$. In the special case where the prior can perfectly estimate U_J from $W_0, W_1, W_2, \dots, W_{t-1}$, we can set $P_t = Q_t$ and $\text{KL}(Q_t \parallel P_t)$ will be zero. As will be seen in the next subsection, we can explicitly design a prior that uses the information in the optimization trajectory for the LD algorithm.

The process by which the prior can learn from the trajectory can be viewed as an *online hypothesis test*, or binary decision problem, where the prior at time t allocates belief between 2^m possible explanations, given by the possible values of U_J , based on the evidence provided by W_0, \dots, W_t . If the prior is able to identify U_J based on the W s then the bound stops accumulating, even if the gradients taken by subsequent training steps are large. This means that penalties for information obtained later in training are *discounted* based on the information obtained earlier in training.

3.4.2 Example: Langevin Dynamics Algorithm for Non-Convex Learning

We apply these results to obtain generalization bounds for a gradient-based iterative noisy algorithm, the Langevin Dynamics (LD) algorithm. For classification with continuous parameters, the 0-1 loss does not provide useful gradients. Typically we optimize a surrogate objective, based on a *surrogate* loss, such as cross entropy. Write $\tilde{\ell} : \mathcal{Z} \times \mathcal{W} \rightarrow \mathbb{R}$ for the surrogate loss and let $\tilde{R}_S(w) = \frac{1}{n} \sum_{i=1}^n \tilde{\ell}(Z_i, w)$ be the empirical

surrogate risk. Let η_t be the learning rate at time t , β_t the inverse temperature at time t and let ϵ_t be sampled i.i.d. from $\mathcal{N}(0, \mathbb{I}_d)$. Then the LD algorithm iterates are given by

$$W_{t+1} = W_t - \eta_t \nabla \tilde{R}_S(W_t) + \sqrt{\frac{2\eta_t}{\beta_t}} \epsilon_t. \quad (3.11)$$

The prior We will take $m = 1$, and construct a bespoke $\sigma(Z^{(2)}, U_{J^c}, J)$ -measurable prior for this problem in order to apply Theorem 3.3.7. The prior is based on a *decision function*, $\theta : \mathbb{R} \rightarrow [0, 1]$, which at each time $t + 1$ takes in a $\sigma(W_0 \dots W_t)$ -measurable *test statistic*, ΔY_t , and returns a *degree of belief* in favor of the hypothesis $U_J = 1$ over $U_J = 2$. The prior predicts an LD step by replacing the unknown (to the prior) contribution to the gradient of the data point at index J with a $\hat{\theta}_t = \theta(\Delta Y_t)$ -weighted average of the gradients due to each candidate $\{Z_{1,J}, Z_{2,J}\}$. The conditional law of the t th iterate under the prior is a $\sigma(Z^{(2)}, U_{J^c}, J, W_0, \dots, W_t)$ -measurable random measure, as required. The exact value of the test statistic is $\Delta Y_t = Y_{t,2} - Y_{t,1}$, here the $Y_{0,1} = Y_{0,2} = 0$ and $Y_{t,u}$ are defined by the formula in Eq. (3.13). The conditional law of the t th iterate under the prior is described by

$$W_{t+1} = W_t - \frac{\eta_t}{n} \left(\sum_{\substack{i=1 \\ i \neq J}}^n \nabla \tilde{\ell}(Z_i, W_t) + \hat{\theta}_t \nabla \tilde{\ell}(Z_{1,J}, W_t) + (1 - \hat{\theta}_t) \nabla \tilde{\ell}(Z_{2,J}, W_t) \right) + \sqrt{\frac{2\eta_t}{\beta_t}} \epsilon_t. \quad (3.12)$$

The test statistic chosen is based on the log-likelihood-ratio test statistic for the independent mean 0 Gaussian random vectors $(\epsilon_s)_{s=1}^t$, which is well known to be *uniformly most powerful* for the binary discrimination of means. Natural choices for θ are symmetric CDFs, since they treat possible values of U symmetrically, and are monotone in the test statistic.

We define the *two-sample incoherence* at time t by $\zeta_t = \nabla \tilde{\ell}(Z_{1,J}, W_t) - \nabla \tilde{\ell}(Z_{2,J}, W_t)$. Θ denotes the set of measurable $\theta : \mathbb{R} \rightarrow [0, 1]$. $Y_{0,1} = Y_{0,2} = 0$, and for $t \geq 1$, $Y_{t,1}$ and $Y_{t,2}$ are given by (for $u \in \{1, 2\}$)

$$Y_{t,u} \triangleq \sum_{i=1}^t \frac{\beta_{i-1}}{4\eta_{i-1}} \left\| W_i - W_{i-1} + \eta_{i-1} \frac{n-1}{n} \nabla \tilde{R}_{S_{J^c}}(W_{i-1}) + \frac{\eta_{i-1}}{n} \nabla \tilde{\ell}(Z_{u,J}, W_{i-1}) \right\|^2. \quad (3.13)$$

Theorem 3.4.2 (Generalization bound for LD algorithm). *Let $\{W_t\}_{t \in [T]}$ denote the*

iterates of the LD algorithm. If $\ell(Z, w)$ is $[0, 1]$ -bounded then

$$\mathbb{E} \left[R_{\mathcal{D}}(W_T) - \hat{R}_S(W_T) \right] \leq \frac{1}{n\sqrt{2}} \inf_{\theta \in \Theta} \mathbb{E} \sqrt{\sum_{t=0}^{T-1} \mathbb{E}^{Z^{(2)}, U, J} \beta_t \eta_t \|\zeta_t\|^2 \left(\mathbb{1}\{U_J = 1\} - \theta(Y_{t,2} - Y_{t,1}) \right)^2}. \quad (3.14)$$

Remark 3.4.3. For $\theta \in \Theta$ with $1 - \theta(x) = \theta(-x)$, Eq. (3.14) simplifies to

$$\mathbb{E} \left[R_{\mathcal{D}}(W_T) - \hat{R}_S(W_T) \right] \leq \frac{1}{n\sqrt{2}} \mathbb{E} \sqrt{\sum_{t=0}^{T-1} \mathbb{E}^{Z^{(2)}, U, J} \beta_t \eta_t \|\zeta_t\|^2 \theta^2(-1^{U_J} (Y_{t,2} - Y_{t,1}))}. \quad (3.15)$$

For instance $\theta(x) = \frac{1}{2} + \frac{1}{2} \tanh(x)$ and $\theta(x) = \frac{1}{2} + \frac{1}{2} \text{sign}(x)$ satisfy $1 - \theta(x) = \theta(-x)$. \triangleleft

Remark 3.4.4. By the law of total expectation, for any $\theta \in \Theta$, $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{1}{2\sqrt{2n}} \mathbb{E} [V_1 + V_2]$, where

$$V_u \triangleq \sqrt{\sum_{t=0}^{T-1} \mathbb{E}^{Z^{(2)}, U_{J^c}, J, U_J=u} \beta_t \eta_t \|\zeta_t\|^2 \left(\mathbb{1}\{u = 1\} - \theta(Y_{t,2} - Y_{t,1}) \right)^2}, \quad u \in \{1, 2\}. \quad (3.16)$$

To estimate V_u ($u \in \{1, 2\}$) for fixed J , the training set is $S_u = \{Z_1, \dots, Z_{J-1}, \tilde{Z}_{u,J}, Z_{J+1}, \dots, Z_n\}$. Hence V_1, V_2 can be simulated from just $n+1$ data points $(Z_1, \dots, Z_{J-1}, Z_{J+1}, \dots, Z_n, \tilde{Z}_{1,J}, \tilde{Z}_{2,J}) \sim \mathcal{D}^{\otimes (n+1)}$. \triangleleft

The generalization bound in Eq. (3.14) does not place any restrictions on the learning rate or Lipschitz continuity of the loss or its gradient. In the next corollary we study the asymptotic properties of the bound in Eq. (3.14) when $\tilde{\ell}$ is L -Lipschitz. Then, we draw a comparison between the bound in this chapter and some of the existing bounds in the literature.

Corollary 3.4.5. *Under the assumption that $\tilde{\ell}$ is L -Lipschitz, we have $\|\zeta_t\| \leq 2L$. Then, the generalization bound in Eq. (3.14) can be upper-bounded as*

$$\mathbb{E}(R_{\mathcal{D}}(W_T) - R_S(W_T)) \leq \frac{\sqrt{2}L}{n} \inf_{\theta \in \Theta} \mathbb{E} \sqrt{\sum_{t=0}^{T-1} \mathbb{E}^{Z^{(2)}, U, J} \beta_t \eta_t \left(\mathbb{1}\{U_J = 1\} - \theta(Y_{t,2} - Y_{t,1}) \right)^2}. \quad (3.17)$$

Remark 3.4.6. Under an L -Lipschitz assumption, for the LD algorithm, J. Li, Luo,

and Qiao [LLQ20, Thm. 9] have

$$\mathbb{E} [R_{\mathcal{D}}(W_T) - R_S(W_T)] \leq \frac{\sqrt{2}L}{n} \sqrt{\sum_{t=0}^{T-1} \beta_t \eta_t}. \quad (3.18)$$

We immediately see that Eq. (3.17) provides a constant factor improvement over Eq. (3.18) by naively using $\theta : x \mapsto 1/2$. Our bound has order-wise improvement with respect to n over that of Bu, S. Zou, and Veeravalli [BZV20a] and Pensia, Jog, and Loh [PJL18] under the L -Lipschitz assumption. Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19, App. E.1] obtain

$$\mathbb{E} [R_{\mathcal{D}}(W_T) - R_S(W_T)] \leq \frac{L}{2(n-1)} \sqrt{\sum_{t=0}^{T-1} \beta_t \eta_t}. \quad (3.19)$$

which is a constant factor better than our bound for the choice $\theta : x \mapsto 1/2$. However, this θ essentially corresponds to no hypothesis test, yielding the same prior as in [NHDKR19]. For more sophisticated choices of decision function (θ), even under a Lipschitz-surrogate loss assumption, it is difficult to compare our bound with related work because the exact impact of θ -discounting is difficult to quantify analytically. \triangleleft

Remark 3.4.7. A prevailing method for analyzing the generalization error in [NHDKR19; BZV20a; PJL18; LLQ20] for iterative algorithms is via the chain rule for KL, using priors for the joint distribution of weight vectors that are Markov, i.e., given the t th weight, the $(t + 1)$ th weight is conditionally independent from the trajectory so far. Existing results using this approach accumulate a "penalty" for each step. In [NHDKR19; BZV20a; LLQ20], the penalty terms are, respectively, the squared Lipschitz constant, the squared norm of the gradients, and the trace of the minibatch gradient covariance. The penalty term in our bound is the squared norm of "two-sample incoherence", defined in Theorem 3.4.2 as the squared norm of the difference between the gradient of a randomly selected training point and the held-out point. However, the use of chain rule along with existing "Markovian" priors introduces a source of looseness, i.e., the accumulating penalty may diverge to $+\infty$ yielding vacuous bounds (as seen in Fig. 1). *The distinguishing feature of our data-dependent CMI analysis is that the penalty terms get "filtered" by the online hypothesis test via our non-Markovian prior, i.e., our prediction for $t + 1$ depends on whole trajectory.* When the true index can be inferred from the previous weights, then the penalty essentially stops accumulating. \triangleleft

Empirical Results

In order to better understand the effect of discounting and the degree of improvement due to our new bounds and more sophisticated prior, we turn to simulation studies. We present and compare the empirical evaluations of the generalization bound in Theorem 3.4.2 with the data-dependent generalization bounds in Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19] and J. Li, Luo, and Qiao [LLQ20]. For brevity, many of the details behind our implementation are deferred to Appendix B.7. The functional form of our bounds and [NHDKR19; LLQ20] are nearly identical as all of them use the chain rule for KL divergence. Nevertheless, the summands appearing in the bounds are different. The bound in [LLQ20] depends on the squared surrogate loss gradients norm, and the generalization bound in Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19] depends on the squared norm of *training set incoherence* defined as $\|\nabla\tilde{\ell}(Z_J, W_t) - \frac{1}{n-1} \sum_{i \in [n], i \neq J} \nabla\tilde{\ell}(Z_i, W_t)\|^2$ where the training set is $\{Z_1, \dots, Z_n\}$ and $J \sim \text{Unif}([n])$. The first key difference between our bound and others is that the summand in our bound consists of two terms: squared norm of the two-sample incoherence, i.e., $\|\zeta_t\|^2$, and the squared error probability of a hypothesis test at time t , given by the term $(\mathbf{1}\{U_J = 1\} - \theta(\sum_{i=0}^t (Y_{i,2} - Y_{i,1})))^2$ in our bound. A consequence of this, and the second fundamental difference between our bound and existing bounds, is that our bound exhibits a trade-off in $\|\zeta_t\|^2$ because large $\|\zeta_t\|^2$ will make the error of the hypothesis test small on future iterations, whereas the bounds in [NHDKR19; LLQ20] are uniformly increasing with respect to the squared norm of surrogate loss gradients and the training set incoherence, respectively. In this section we empirically evaluate and compare our bound with related work across various neural network architectures and datasets.

Using Monte Carlo (MC) simulation, we compared estimates of our expected generalization error bounds with estimates of the bound from [NHDKR19; LLQ20] for the MNIST [LCB10], CIFAR10 [Kri09], and Fashion-MNIST [XRV17] datasets in Fig. 3.1 and Table 3.1. For all the plots we consider $\theta(x) = \frac{1}{2}(1 + \text{erf}(x))$ for our bound. Also, in the last row of Table 3.1, we report the *unbiased estimate* of our bound optimized over the choice of θ function. We plot the squared norm of the two sample incoherence and training set incoherence, as well as the squared error probability of the hypothesis test. Fig. 3.1 and Table 3.1 show that our bound is tighter, and remain non-vacuous after many more iterations. We also observe that the variances for MC estimates of our bound are smaller than those of Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19], and it is also smaller than J. Li, Luo, and Qiao [LLQ20] for CIFAR10 and MNIST-CNN experiments. Moreover, we observe that

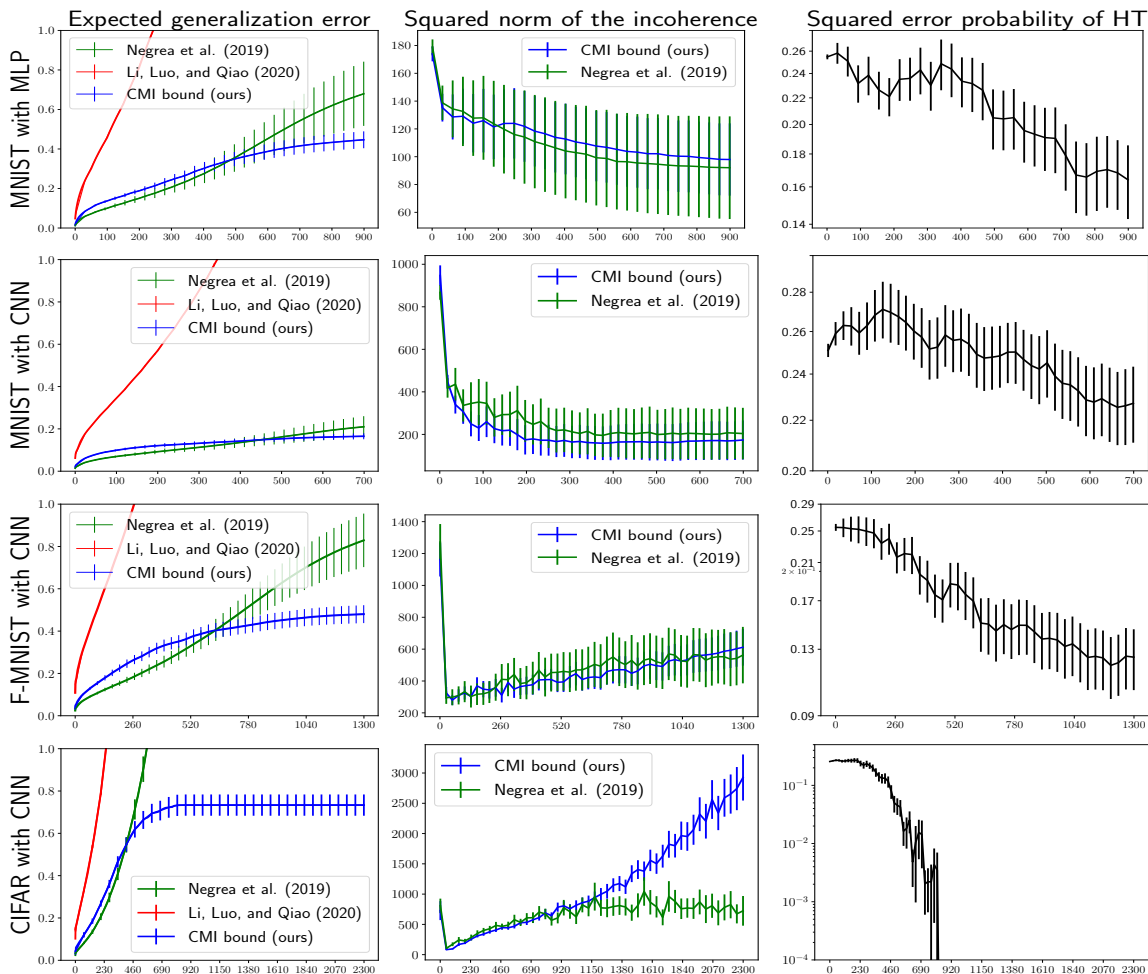


Figure 3.1: Numerical results for various datasets and architectures. All the x-axes represent the training iteration. The plots in the first column depict a Monte Carlo estimate of our bounds with that of Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19] and J. Li, Luo, and Qiao [LLQ20]. The plots in the second column compare the mean of the *training set incoherence* in [NHDKR19] with the two-sample incoherence in our bound. Finally, the plots in the third column show the mean of the squared error probability of the hypothesis testing performed by the proposed prior in our bound.

	MNIST-MLP	MNIST-CNN	CIFAR10-CNN	FMNIST-CNN
Training error	4.33 ± 0.01%	2.59 ± 0.01%	9.39 ± 0.36%	7.96 ± 0.03%
Generalization error	0.88 ± 0.01%	0.55 ± 0.01%	32.89 ± 0.44%	3.71 ± 0.03%
Our(NeurIPS'19)	67.93 ± 16.25%	20.98 ± 5.01%	4112.63 ± 567.08%	82.89 ± 12.64%
J. Li, Luo, and Qiao [LLQ20]	600.29 ± 1.99%	245.03 ± 2.37%	20754.32 ± 75.95%	598.62 ± 3.21%
Our(NeurIPS'20)	44.65 ± 4.27%	16.51 ± 1.41%	71.76 ± 4.82%	48.01 ± 4.22%
Our-Optimized(NeurIPS'20)	39.06 ± 5.52%	13.24 ± 1.53%	63.00 ± 5.97%	41.17 ± 5.85%

Table 3.1: Summary of the results. The generalization bounds are reported at the end of training.

the error probability of the hypothesis test decays with the number of iterations, which matches the intuition that, as one observes more noisy increments of the process, one is more able to determine which point is contributing to the gradient. For CIFAR10, $\|\zeta_t\|^2$ is large because the generalization gap is large. However, as mentioned in the beginning of this section, large $\|\zeta_t\|^2$ makes the hypothesis testing easier on subsequent iterations. For instance, after iteration 600 the error is vanishingly small for CIFAR10 experiments which results in a plateau region in the bound. We can also observe the same phenomenon for the Fashion-MNIST experiment. This property distinguishes our bound from those in [NHDKR19; LLQ20].

Results for MNIST with CNN demonstrate that $\|\zeta_t\|^2$ and training set incoherence are close to each other. The reason behind this observations is that the generalization gap is small. Also, for this experiment the performance of the hypothesis testing is only slightly better than random guessing since the generalization gap is small, and it is difficult to distinguish the training samples from the test samples. This observation explains why our generalization bound is close to that of [NHDKR19]. Nevertheless, the hypothesis testing performance improves with more training iterations, leading the two bounds to diverge, with our new bound performing better at later iterations. Finally, the scaling of our bound with respect to the number of iteration is tighter than in the bounds in [NHDKR19; LLQ20] as can be seen in Fig. 3.1.

Chapter 4

Towards a Unified Information-Theoretic Framework for Generalization

4.1 Introduction

In this chapter, we study the expressiveness of generalization bounds in terms of information-theoretic measures of dependence between the output of a learning algorithm and input data. Let \mathcal{D} be an unknown distribution on a space \mathcal{Z} , and let \mathcal{H} be a set of classifiers. Consider a (randomized) learning algorithm $\mathcal{A} = (\mathcal{A}_n)_{n \geq 1}$ that selects an element \hat{h} in \mathcal{H} , based on i.i.d. samples $S_n \sim \mathcal{D}^{\otimes n}$, i.e., $\hat{h} = \mathcal{A}_n(S_n)$. As stated above, the initial focus of this line of work was on the *input-output mutual information (IOMI)* of an algorithm $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n) = I(\mathcal{A}_n(S_n); S_n)$ between the input and the output of a learning problem. A natural question is whether IOMI framework can provide a sharp characterization of the learnability of Vapnik–Chervonenkis (VC) classes, for which we have strong generalization guarantees.

A negative resolution was provided by Bassily, Moran, Nachum, Shafer, and Yehudayoff [BMNSY18] for the concept class of thresholds in one dimension. Follow up work by Nachum, Shafer, and Yehudayoff [NSY18] extended the argument in [BMNSY18], proving the following result:

Theorem 4.1.1 (Thm. 1, [NSY18]). *For every $d \in \mathbb{N}$ and every $n \geq 2d^2$, there exists a finite input space \mathcal{X} and a concept class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ of VC-dimension d such that, for all proper and consistent learning algorithm \mathcal{A}_n , there exists a realizable distribution \mathcal{D} such that $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n) = \Omega(d \log \log(|\mathcal{X}|/d))$.*

Livni and Moran [LM20] extended this result even further, showing that, for the class of one-dimensional thresholds over $\{1, \dots, m\}$, $m \in \mathbb{N}$,¹ for *every* learning algorithm \mathcal{A} there exists a realizable distribution such that either the risk (population loss) is large or the $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ scales with the cardinality of the space, m . These results highlight an important limitation of the IOMI framework: given an unbounded input space, for any “good” learning algorithm there are always scenarios in which $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ is unbounded. Therefore, the distribution-free learnability of VC classes cannot be expressed using the IOMI framework.

In this paper, we focus on the “*conditional mutual information*” (CMI) framework, proposed by Steinke and Zakyntinou [SZ20a]. In order to reason about the generalization error of a learning algorithm, they introduce a super sample that contains the training sample as a random subset and compute the mutual information between the input and output conditional on the super sample (formal definitions are provided in Section 4.2.1). Improvements of this framework and its application in studying the generalization of specific learning algorithms have been studied in [HGKS20; HNKRD20; HD20; RBTS20; ZTL21; HD21].

The current paper revolves around the following fundamental question: For which learning problems and learning algorithms is the CMI framework expressive enough to accurately estimate the generalization error? We will focus in particular on whether we can recover optimal worst case (minimax) rates for VC classes satisfying certain properties. The answer to these question provide evidence that the CMI framework provides a unifying framework for studying generalization.

For VC classes, Steinke and Zakyntinou [SZ20a] revealed a stark separation between the CMI framework and IOMI framework. They showed the existence of an empirical risk minimization (ERM) algorithm whose CMI is no larger than $d \log n + 2$ for learning every VC class of dimension d given n i.i.d. training samples. In contrast to Theorem 4.1.1, CMI does not scale with the cardinality of the space. However, the bound on the CMI combined with Steinke and Zakyntinou’s CMI-based generalization bound, leads to a bound on the expected excess risk that is suboptimal in some cases by a $\log n$ factor. (For an overview of the known bounds for learning VC classes, please refer to Appendix C.1.) The suboptimality of their bound prompted Steinke and Zakyntinou [SZ20b] to conjecture that the CMI bound for proper learners of VC classes can be improved to $O(d)$. Moreover, Steinke and Zakyntinou connected

¹ This concept class can be defined as follows. Let $\mathcal{X} = \{1, \dots, m\}$. Let $k \in \mathbb{N}$ and $h_k : \mathcal{X} \rightarrow \{0, 1\}$ define as $h_k(x) = \mathbb{1}[x > k]$. Then, the class of one-dimensional thresholds over $\{1, \dots, m\}$ is $\mathcal{H}_m = \{h_k : \mathcal{X} \rightarrow \{0, 1\} | k \in \mathbb{N}\}$.

CMI framework to the sample compression framework of [LW86] by showing that a sample compression \mathcal{A}_n of size k has $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq k \log 2n$. Their bound for sample compression schemes is also suboptimal in some cases by a $\log n$ factor.

4.1.1 Contributions

In this paper we extend the reach of the CMI framework by demonstrating its unifying nature for obtaining optimal or near-optimal bounds for the expected excess risk of the various algorithms in the realizable setting.

1. We demonstrate that one can use the CMI framework to express non-trivial (but sub-optimal) bounds for every improper learning algorithm that outputs a hypothesis from a class with a bounded VC dimension. This is achieved by an empirical variant of CMI defined by [SZ20a].
2. We study the CMI of SVMs for learning half spaces and show that the CMI framework yields optimal bounds on the expected excess risk. Our bound on the CMI of SVM is an application of our general result giving optimal CMI bounds for stable sample compression schemes [BHMZ20; HK21], which improve on CMI bounds for general sample compression schemes [SZ20a] by a $\log n$ factor.
3. In the context of proper learning of VC classes, we exhibit VC classes for which the CMI of any proper learner cannot be bounded by any real-valued function of the VC dimension. Then, we consider VC classes with finite star number [HY15], and prove the existence of a learner with bounded CMI. Finally, we show that the release of the set of all consistent classifiers in \mathcal{H} has bounded CMI *if and only if* \mathcal{H} has finite star number.
4. We show that CMI framework is *universal* in the realizable setting. More precisely, for every data distribution and consistent learner, the bound on excess risk obtained by the CMI framework vanishes if and only if the excess risk also vanishes as the number training samples diverges. We then show that any learning algorithm with a “leave-one-out” bound of order $O(1/n)$ yields an evaluated-CMI bound of order $O(\log n)$. As an application, we study the classical one-inclusion graph algorithm of Haussler, Littlestone, and M. K. Warmuth [HLW94] for improper learning of VC classes, and provide a nearly optimal bound on its expected excess risk using the CMI framework. We also prove there exists a randomized one-inclusion graph which learns point functions (singleton) with bounded CMI.

Our results indicate that CMI is a very expressive generalization framework, and one that can tie together existing frameworks. Although most of our results are stated for binary classification in the distribution-free setting, it is interesting to note that the CMI framework is known to provide numerically non-vacuous generalization error guarantees for some modern deep learning models and datasets in the distribution-dependent setting [HNKRD20; HD21]. These developments in a range of different problem settings highlight the importance of understanding the expressiveness of the CMI framework.

4.2 Preliminaries

We consider the problem of binary classification, with inputs in some space \mathcal{X} assigned labels in $\mathcal{Y} = \{0, 1\}$. A concept (or hypothesis) class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a set of functions $h : \mathcal{X} \rightarrow \mathcal{Y}$. We say \mathcal{H} *shatters* $(x_1, \dots, x_m) \in \mathcal{X}^m$ if for all $(y_1, \dots, y_m) \in \{0, 1\}^m$, there exists $h \in \mathcal{H}$, such that, for all $i \in [m]$, we have $h(x_i) = y_i$. The *VC dimension* of \mathcal{H} , denoted by d , is the largest $m \in \mathbb{N}$ for which there exists $(x_1, \dots, x_m) \in \mathcal{X}^m$ shattered by \mathcal{H} . If no such finite m exists, then $d = \infty$.

Let \mathcal{D} be a distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The *empirical (classification) risk* of a classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ on a sample $s = ((x_1, y_1), \dots, (x_n, y_n)) \in \mathcal{Z}^n$ is $\hat{R}_s(h) = n^{-1} \sum_{i \in [n]} \ell(h, (x_i, y_i))$, where $\ell(h, (x, y)) = \mathbb{1}[h(x) \neq y]$. Let $S_n \sim \mathcal{D}^n$, i.e., let S_n be a sequence of i.i.d. random elements in \mathcal{Z} with common distribution \mathcal{D} . (We can view S_n itself as a random element in \mathcal{Z}^n .) The *risk* of h is $R_{\mathcal{D}}(h) = \mathbb{E} \hat{R}_{S_n}(h)$, where \mathbb{E} denotes the expectation operator. (The risk has, of course, no dependence on n due to the data being i.i.d.)

A distribution \mathcal{D} is *realizable by a class* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists $h \in \mathcal{H}$ such that $R_{\mathcal{D}}(h) = 0$. A sequence $((x_1, y_1), \dots, (x_n, y_n))$ is said to be *realizable by* \mathcal{H} , if for some $h \in \mathcal{H}$, $h(x_i) = y_i$ for all $i \in [n] = \{1, \dots, n\}$. Note that if a distribution is realizable by \mathcal{H} , it implies that with probability one over $S_n \sim \mathcal{D}^n$, the training sample S_n is realizable by \mathcal{H} .

Let $\mathcal{A} = (\mathcal{A}_n)_{n \geq 1}$ denote a (potentially randomized) learning algorithm, which, for any positive integer n , maps S_n to an element of $\mathcal{X} \rightarrow \mathcal{Y}$. We say that \mathcal{A} is a *proper learner for a class* $\mathcal{H} \subseteq \mathcal{X} \rightarrow \mathcal{Y}$ if the codomain of \mathcal{A}_n is a subset of \mathcal{H} for every n . We say \mathcal{A}_n is a *consistent algorithm (learner)* if $\hat{R}_{S_n}(\mathcal{A}_n(S_n)) = 0$ a.s. Our primary interest in this paper is the *expected generalization error* of \mathcal{A}_n with respect to \mathcal{D} , defined as $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) = \mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n)) - \hat{R}_{S_n}(\mathcal{A}_n(S_n))]$, where we average over both the choice of training sample and the randomness within the algorithm \mathcal{A}_n .

4.2.1 Conditional mutual information (CMI) of an algorithm

In order to study generalization, and avoid some of the pitfalls of earlier approaches based on mutual information, Steinke and Zakynthinou [SZ20a] propose to study the information contained in a “supersample” Z , a training sample S_n taken from the supersample, and the hypothesis $\mathcal{A}_n(S_n)$ output by a possibly randomized learning algorithm, given S_n as input. Formally, let $Z = (Z_{i,j})_{i \in \{0,1\}, j \in [n]}$ to be an array of i.i.d. random elements in the space \mathcal{Z} of labeled examples, with a common distribution \mathcal{D} . In order to choose a training sample S_n of size n from Z , let $U = (U_1, U_2, \dots, U_n)$ be a sequence of i.i.d. Bernoulli random variables in $\{0, 1\}$, independent from Z , with $\mathbb{P}(U_i = 0) = 1/2$. Define $S_n = Z_U = (Z_{U_j, j})_{j=1}^n$. The *conditional mutual information (CMI) of \mathcal{A}_n* , denoted $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$, is defined to be the mutual information between $\mathcal{A}_n(S_n)$ and U given Z , denoted $I(\mathcal{A}_n(S_n); U|Z)$. This quantity is equivalent to $I(\mathcal{A}_n(S_n); S_n|Z)$ when \mathcal{D} is atomless, since (U_1, \dots, U_n) is a.s. measurable with respect to S_n and Z . Because Z and U are independent, $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq H(U|Z) = H(U) = n \log 2$. We now pause to introduce this and other information-theoretic quantities formally.

4.2.2 Measures of divergence and information

Let P, Q be probability measures on a measurable space. (We ignore measure-theoretic pathologies for clarity.) For a P -integrable or nonnegative function f , let $P[f] = \int f dP$. When Q is absolutely continuous with respect to P , denoted $Q \ll P$, write $\frac{dQ}{dP}$ for (an arbitrary version of) the Radon–Nikodym derivative (or density) of Q with respect to P . The *KL divergence (or relative entropy) of Q with respect to P* , denoted $\text{KL}(Q \| P)$, is defined as $Q[\log \frac{dQ}{dP}]$ when $Q \ll P$ and infinity otherwise.

For a random element X in some measurable space \mathcal{X} , let $\mathbb{P}[X]$ denote its distribution, which lives in the space $\mathcal{M}_1(\mathcal{X})$ of all probability measures on \mathcal{X} . Given another random element, say Y in \mathcal{T} , let $\mathbb{P}^Y[X]$ denote the conditional distribution of X given Y . If X and Y are independent, $\mathbb{P}^Y[X] = \mathbb{P}[X]$ a.s. For an event, say $X \in A$, $\mathbb{P}^Y[X \in A]$ denotes the event’s conditional probability given Y , which is defined to be the conditional expectation of the indicator random variable $\mathbf{1}[X \in A]$ given Y , denoted $\mathbb{E}^Y \mathbf{1}[X \in A]$.² By the chain (aka tower) rule, $\mathbb{E}\mathbb{E}^{\mathcal{F}} = \mathbb{E}$ for any σ -algebra \mathcal{F} .

The *mutual information between X and Y* is $I(X; Y) = \text{KL}(\mathbb{P}[(X, Y)] \| \mathbb{P}[X] \otimes$

² By definition, $\mathbb{P}^Y[X]$ is a $\sigma(Y)$ -measurable random element in $\mathcal{M}_1(\mathcal{X})$, i.e., $\mathbb{P}^Z[U] = \kappa(Z)$ a.s. for some measurable map $\kappa : \mathcal{T} \rightarrow \mathcal{M}_1(\mathcal{X})$. More generally, if, say $\mathcal{F} = \sigma(Y, Z)$ is the σ -algebra generated by Y and Z , then a conditional distribution/probability/expectation given \mathcal{F} is a measurable function of Y and Z .

$\mathbb{P}[Y]$), where \otimes forms the product measure. Writing $\mathbb{P}^Z[(X, Y)]$ for the conditional distribution of the pair (X, Y) given a random element Z , the *disintegrated mutual information between X and Y given Z* , is

$$I^Z(X; Y) = \text{KL}(\mathbb{P}^Z[(X, Y)] \parallel \mathbb{P}^Z[X] \otimes \mathbb{P}^Z[Y]).$$

Then the *conditional mutual information* of X and Y given Z is $I(X, Y|Z) = \mathbb{E}I^Z(X, Y)$.

Let $\mu = \mathbb{P}[X]$ and let $\kappa(Y) = \mathbb{P}^Y[X]$ a.s. If X concentrates on a countable set V with counting measure ν , the (*Shannon*) *entropy of X* is $H(X) = -\mu[\log \frac{d\mu}{d\nu}] = -\sum_{x \in V} \mathbb{P}(X = x) \log \mathbb{P}(X = x)$. The *disintegrated entropy of X given Y* is defined by $H^Y(X) = -\kappa(Y)[\log \frac{d\kappa(Y)}{d\nu}]$, while the *conditional entropy of X given Y* is $H(X|Y) = \mathbb{E}[H^Y(X)]$. Note that $H(X|Y) \leq H(X)$. We will make use of the following lemma whose proof can be found in [MT07].

Lemma 4.2.1. *Let (X_1, X_2, \dots, X_n) be a discrete random vector, and Y be an arbitrary random variable. Then, $H(X_1, \dots, X_n|Y) \geq \sum_{i=1}^n H(X_i|X_{-i}, Y)$, where $X_{-i} = (X_j : j \in [n], j \neq i)$.*

Steinke and Zakyntinou establish a range of generalization bounds in terms of CMI. Our primary interest is in bounds for algorithms that have vanishing empirical risk. For $[0, 1]$ -bounded loss, Steinke and Zakyntinou show that

$$\mathbb{E}R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \leq 2\mathbb{E}\hat{R}_S(\mathcal{A}_n(S_n)) + \frac{3\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}. \quad (4.1)$$

For consistent learners (i.e., those that achieve zero empirical error a.s.), they also establish

$$\mathbb{E}R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \leq \frac{\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n \log 2}. \quad (4.2)$$

Steinke and Zakyntinou also introduce a variant of CMI based on the information revealed by the learner's losses on Z , rather than by the output hypothesis, $\mathcal{A}_n(S_n)$, directly.

Definition 4.2.2 (Evaluated CMI, [SZ20a, §6.2.2]). Let $L \in \{0, 1\}^{2 \times n}$ be the array with entries $L_{i,j} = \ell(\mathcal{A}_n(S_n), Z_{i,j})$ for $i \in \{0, 1\}$, $j \in [n]$. The *evaluated conditional mutual information of \mathcal{A}_n with respect to \mathcal{D}* , denoted by $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))$, is the conditional mutual information $I(L; U|Z)$.

By the data processing inequality, $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) \leq \text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$. Therefore, $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))$ is also bounded above by $n \log 2$. For consistent learners, Steinke and Zakyntinou show

$$\mathbb{E}R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \leq 1.5 \frac{\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))}{n}. \quad (4.3)$$

For consistent learners \mathcal{A}_n with bounded CMI or eCMI, these results imply their expected excess risk is of order $O(1/n)$. The following result gives a nearly optimal bound for the generalization error for VC classes in term of the evaluated CMI. The proof (Appendix C.2) uses standard arguments, controlling the cardinality of the support of L using the Sauer–Shelah lemma.

Theorem 4.2.3. *For every n , let $\mathcal{A}_n : \mathcal{Z}^n \rightarrow \mathcal{H}_n$, where \mathcal{H}_n is a concept class with VC dimension d_n . Then, for every n and distribution on Z , $I^Z(L; U) \leq d_n \log 6n$ a.s. In particular, $\sup_{\mathcal{D}} \text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) \in O(d_n \log n)$.*

Remark 4.2.4. Markov’s inequality and Eq. (4.2) imply $\mathbb{P}(R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \geq \epsilon) \leq \text{CMI}_{\mathcal{D}}(\mathcal{A}_n)/(\log(2)n\epsilon)$ for consistent learners. By [HNKRD20, Thm. 2.1], $I(\mathcal{A}_n(S_n); U|Z) \leq I(\mathcal{A}_n(S_n); S_n)$. This observation, combined with [BMNSY18, Prop. 11], implies there is an input space, data distribution, and consistent learning algorithm for which this tail bound’s dependence on n is *tight*. If one were to obtain sample complexity bounds via such tail bounds, one would only prove that $O(1/(\epsilon\delta))$ samples suffice to find a hypothesis with ϵ estimation error with probability at least $1 - \delta$. The linear dependence on $1/\delta$ is, however, suboptimal. As such, it seems that the CMI framework cannot be used to obtain optimal sample complexity bounds in the PAC framework. Recent proposals for disintegrated notions of CMI in [HD20] might provide a framework for studying the sample complexity of PAC learning using an information-theoretic framework. \triangleleft

4.3 Optimal CMI Bound for SVM and Stable Compression Schemes

In this section, we show that the CMI framework can be used to derive an optimal excess risk bound for the SVM algorithm learning half spaces in \mathbb{R}^d . To show this, we establish optimal CMI bounds for the subclass of *stable* sample compression schemes, which imply this section’s main result:

Theorem 4.3.1. *Let \mathcal{A}_n be the SVM algorithm for learning the class of half spaces in \mathbb{R}^d . Then, for every $n > d/2$ and realizable distribution \mathcal{D} in \mathbb{R}^d , we have $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 2(d+1)\log 2$.*

Combining this result with Eq. (4.2) gives $\mathbb{E}R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \leq 2(d+1)/n$. The lower bound for expected excess risk of linear classifiers in [LL20] shows this bound is optimal up to a constant factor.

4.3.1 CMI of Stable Compression Schemes

Littlestone and M. Warmuth [LW86] introduced compression schemes, which capture the idea that a consistent hypothesis can be defined in terms of a fixed number of samples. Formally, for a concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, a *sample compression scheme* of size $k \in \mathbb{N}$ is a pair (κ, ρ) of maps such that, for all samples $s = ((x_i, y_i))_{i=1}^n$ of size $n \geq k$, the map κ compresses the sample into a length- k subsequence $\kappa(s) \subseteq s$ which the map ρ uses to reconstruct an empirical risk minimizer $\hat{h} = \rho(\kappa(s))$. Steinke and Zakynthinou prove the following upper bound on the CMI of a sample compression scheme.

Theorem 4.3.2 ([SZ20a, Thm. 4.1]). *Let \mathcal{H} be a hypothesis class that has a sample compression scheme (κ, ρ) of size k . Then, $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq k \log(2n)$ where $\mathcal{A}_n(\cdot) = \rho(\kappa(\cdot))$.*

Note that the bound in Theorem 4.3.2 *cannot* be improved from $O(k \log n)$ to $O(k)$ for *every* sample compression scheme, and so the bound in Theorem 4.3.2 is tight, and cannot be improved without further information about the compression scheme. The proof of optimality stems from the fact that there exists compression schemes of size k and data distributions \mathcal{D} such that there is a lower bound $\mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n)] = \Omega(k \log(n)/n)$ where $\mathcal{A}_n(\cdot) = \rho(\kappa(\cdot))$ [FW95]. Combining this lower bound with Eq. (4.2) proves the optimality of Theorem 4.3.2.

Nevertheless, we can circumvent this lower bound by considering an important subclass of the sample compression schemes. Many natural compression schemes are also *stable* in the sense that removing any training example that was not in the compressed sequence does not alter the resulting classifier. To give a formal definition, we write $s \subseteq s'$ for two sequences s, s' if, under some permutation, s is a subsequence of s' .

Definition 4.3.3 (Stable sample compression scheme; [BHMZ20]). A sample compression scheme (κ, ρ) of size k is said to be *stable* if κ is symmetric (i.e., invariant to

permutation of its input) and, for every realizable sample s of size $n \geq k$, and every sequence s' such that $\kappa(s) \subseteq s' \subseteq s$, we have $\rho(\kappa(s)) = \rho(\kappa(s'))$. Due to the symmetry of κ , we refer to its output as the compression *set*, although the equivalence class of sequence under permutations is the structure of a multiset, not a set.

The concept of a stable compression scheme has its roots in the analysis of the SVM for learning half-spaces in \mathbb{R}^d [VC74], which is the quintessential example of a stable sample compression scheme. For SVMs, the compression (multi)set contains at most $d + 1$ distinct “support vectors” for any given training set. The reconstruction map outputs the max-margin classifier over the set of support vectors. By stability, removing any training example that is not a support vector does not change the resulting classifier [MRT18, Sec. 5.3.2]. In the next theorem, we present a uniform CMI bound over realizable distributions for every stable sample compression scheme. Our bound removes the $\log n$ factor from Theorem 4.3.2 and is *optimal* up to a constant factor in the distribution-free setting.

Theorem 4.3.4. *Let \mathcal{H} be a concept class with a stable compression scheme (κ, ρ) of size k . Then, for every realizable data distribution \mathcal{D} and $n \geq k$, $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 2k \log 2$, where $\mathcal{A}_n = \rho(\kappa(\cdot))$.*

Remark 4.3.5. Steinke and Zakyntinou [SZ20a, Sec. 4.4] propose an algorithm for learning threshold functions (positive rays) in the realizable setting over \mathbb{R} that achieves $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 2 \log 2$. It is interesting to note that their algorithm can be viewed as a stable compression scheme. Specifically, for a realizable training set s , let $x^* = \min\{x \in \mathbb{R} : (x, 1) \in s\}$ if s has any sample with label 1, otherwise let $x^* = \infty$. Then the algorithm proposed by Steinke and Zakyntinou is $\mathcal{A}_n(S_n) = \hat{h}$, where $\hat{h}(x) = \mathbb{1}[x \geq x^*]$. Steinke and Zakyntinou present a bespoke analysis of this special algorithm. It is straightforward to see that the algorithm is a stable compression scheme of size one and the compression map here is symmetric. Therefore, the result of Theorem 4.3.4 gives $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 2 \log 2$. \triangleleft

Proof of Theorem 4.3.4. Let $W = \mathcal{A}_n(Z_U) = \rho(\kappa(Z_U))$ and note that \mathcal{A}_n is deterministic. We have $H(U|Z) = n \log 2$ due to independence of U and Z and the independence of components of U . Then, by the definition of mutual information in terms of entropy, and Lemma 4.2.1,

$$\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) = H(U|Z) - H(U|W, Z) \leq n \log 2 - \sum_{i=1}^n H(U_i|W, Z, U_{-i}). \quad (4.4)$$

Fix $i \in [n]$, and define $U_{i \rightarrow b} \triangleq (U_1, \dots, U_{i-1}, b, U_{i+1}, \dots, U_n)$ for $b \in \{0, 1\}$. Using this notation, we can define two training sets $S_{i \rightarrow b} = Z_{U_{i \rightarrow b}}$ for $b \in \{0, 1\}$. Let \mathcal{F}_i be the σ -algebra $\sigma(W, Z, U_{-i})$ and let E be the event $\rho(\kappa(S_{i \rightarrow 0})) = \rho(\kappa(S_{i \rightarrow 1}))$. Then, by the non-negativity of entropy,

$$\mathbb{H}(U_i | W, Z, U_{-i}) = \mathbb{E}[\mathbb{H}^{\mathcal{F}_i}(U_i)] \geq \mathbb{E}[\mathbb{H}^{\mathcal{F}_i}(U_i) \mathbf{1}[E]]. \quad (4.5)$$

Note that, conditional on the sub- σ -algebra $\mathcal{G}_i = \sigma(Z, U_{-i})$, W takes on at most two values. However, on the event E (or equivalently, conditioning further on the event E , since E is \mathcal{G}_i -measurable), W is now nonrandom because it takes on a single value. It follows that, conditional on \mathcal{G}_i and the event E , W is trivially independent of every random variable, including U_i . Ergo, on E , $\mathbb{E}^{\mathcal{G}_i}[U_i] = \mathbb{E}^{\mathcal{F}_i}[U_i] = \mathbb{P}^{\mathcal{F}_i}[U_i = 1]$. But U_i is independent of \mathcal{G}_i , and so $\mathbb{E}^{\mathcal{G}_i}[U_i] = \mathbb{E}[U_i] = \frac{1}{2}$. Thus, on E , $\mathbb{P}^{\mathcal{F}_i}[U_i = 1] = \frac{1}{2}$ and so $\mathbb{H}^{\mathcal{F}_i}(U_i) = \mathbb{H}_b(\frac{1}{2}) = \log 2$. Therefore,

$$\mathbb{H}(U_i | W, Z, U_{-i}) \geq \log 2 \cdot \mathbb{P}(\rho(\kappa(S_{i \rightarrow 0})) = \rho(\kappa(S_{i \rightarrow 1}))). \quad (4.6)$$

We can bound the probability of E from below using the stability property of the compression scheme. For any $(x_1, \tilde{x}_1, x_2, \dots, x_n) \in \mathcal{X}^{n+1}$ and $h \in \mathcal{H}$, consider two multisets $S = \{(x_1, h(x_1)), (x_2, h(x_2)), \dots, (x_n, h(x_n))\}$ and $\tilde{S} = \{(\tilde{x}_1, h(\tilde{x}_1)), (x_2, h(x_2)), \dots, (x_n, h(x_n))\}$, where S and \tilde{S} differ only in the first element. Define the multiset $S \cup \tilde{S} = \{(x_1, h(x_1)), (\tilde{x}_1, h(\tilde{x}_1)), (x_2, h(x_2)), \dots, (x_n, h(x_n))\}$. We claim that if $(x_1, h(x_1))$ and $(\tilde{x}_1, h(\tilde{x}_1))$ are not the members of the compression set $S \cup \tilde{S}$, then $(x_1, h(x_1))$ and $(\tilde{x}_1, h(\tilde{x}_1))$ are not in the compression set of S and \tilde{S} , respectively. To prove this claim, since $(x_1, h(x_1))$ is not in the compression set $S \cup \tilde{S}$ by the stability of κ , we have $\rho(\kappa(S \cup \tilde{S})) = \rho(\kappa(S \cup \tilde{S} \setminus \{(x_1, h(x_1))\}))$. By the definition of S and \tilde{S} , $S \cup \tilde{S} \setminus \{(x_1, h(x_1))\} = \tilde{S}$. Thus, combining facts that $(x_1, h(x_1))$ is not in the compression set $S \cup \tilde{S}$ and $\kappa(S \cup \tilde{S}) = \kappa(S \cup \tilde{S} \setminus \{(x_1, h(x_1))\}) = \kappa(\tilde{S})$, we obtain $(\tilde{x}_1, h(\tilde{x}_1))$ is not in the compression set \tilde{S} . Similarly, we can prove $(x_1, h(x_1))$ is not a member of the compression set S by switching x_1 with \tilde{x}_1 in the argument. By this argument,

$$\mathbb{P}(\rho(\kappa(S_{i \rightarrow 0})) = \rho(\kappa(S_{i \rightarrow 1}))) \geq \mathbb{P}(Z_{0,i} \notin \kappa(S_{i \rightarrow 0} \cup S_{i \rightarrow 1}) \wedge Z_{1,i} \notin \kappa(S_{i \rightarrow 0} \cup S_{i \rightarrow 1})). \quad (4.7)$$

Recall that the elements of Z are i.i.d., hence exchangeable. Since the size of the

sample compression is k and κ is symmetric, we have

$$\mathbb{P}(Z_{0,i} \notin \kappa(S_{i \rightarrow 0} \cup S_{i \rightarrow 1}) \wedge Z_{1,i} \notin \kappa(S_{i \rightarrow 0} \cup S_{i \rightarrow 1})) \geq \binom{n-1}{k} / \binom{n+1}{k}. \quad (4.8)$$

Combining Eqs. (4.6) to (4.8) yields $H(U_i | W, Z, U_{-i}) \geq \log 2 \cdot \binom{n-1}{k} / \binom{n+1}{k} \geq (1 - 2k/n) \log 2$. Finally, the result follows by substitution of this bound into Eq. (4.4). \square

The result for SVMs (Theorem 4.3.1) follows immediately from Theorem 4.3.4 and the fact that the SVM may be expressed as a stable compression scheme of size $d + 1$.

4.4 CMI of Proper Learning of VC classes

Following their paper introducing CMI, Steinke and Zakynthinou posed several open problems asking whether VC classes under realizability admit learners with bounded CMI. We will restate their conjectures, and then showing that there exist some VC classes for which it is not possible to find a proper learner with bounded CMI under realizability. We then consider a subset of VC classes, namely VC classes with finite star number, and show that for such concept classes, there exists an ERM with bounded CMI.

We first state the main result of [SZ20a] on the CMI of proper learners.

Theorem 4.4.1 (Thm. 4.12, [SZ20a]). *Let \mathcal{H} be a concept class with VC dimension d . Then for all $n \in \mathbb{N}$, there exists a proper ERM algorithm \mathcal{A}_n for learning \mathcal{H} such that for every realizable distribution \mathcal{D} , $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) = O(d \log n)$.*

Remark 4.4.2 (Comparison of Theorem 4.4.1 and Theorem 4.2.3). First, note that Theorem 4.4.1 does not hold for every ERM algorithm. As discussed in [SZ20a], we can construct pathological ERMs with nearly maximal CMI by simply encoding the information U into the “lower-order” bits of W .

It is also worth noting that our result in Theorem 4.2.3 is more general. There we show that a bound $O(d \log n)$ holds for *evaluated* CMI of *any* algorithm that outputs a hypothesis from VC class, whereas Theorem 4.4.1 holds for a *specific* proper algorithm. \triangleleft

4.4.1 A Limitation of Proper Learning

Steinke and Zakynthinou [SZ20b] propose two conjectures regarding CMI for proper learning of VC classes under the realizability assumption, both of which can be seen as special cases of the following statement:

Statement 1. *There exists a real-valued function f and constant $c \geq 0$ such that, for every nonnegative integer d and VC class $\mathcal{H} \subseteq \mathcal{X} \rightarrow \mathcal{Y}$ of dimension d , there exists a proper learning algorithm \mathcal{A} for \mathcal{H} such that, for every $n \geq d$, $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq f(d)$ for all \mathcal{D} and, for every realizable $s \in \mathcal{Z}^n$, $\mathbb{E}\hat{R}_s(\mathcal{A}_n(s)) \leq cd/n$, where the expectation is taken only over the randomness in \mathcal{A}_n .*

Steinke and Zakyntinou [SZ20b] conjecture that Statement 1 holds for f linear. In this section, we show that Statement 1 is false in general: it is not possible to find a proper learning algorithm for *every* VC class that removes the $\log(n)$ factor from Theorem 4.4.1. For a class $\mathcal{H} \subseteq \mathcal{X} \rightarrow \mathcal{Y}$, let $\mathcal{M}_{\text{prop}}^{\mathcal{H}}(\epsilon, \delta)$ denote the *proper optimal sample complexity* of (ϵ, δ) -PAC learning \mathcal{H} , i.e., $\mathcal{M}_{\text{prop}}^{\mathcal{H}}(\epsilon, \delta)$ is the least integer n , for which there exists a proper learning algorithm \mathcal{A} such that, for every realizable distribution \mathcal{D} , $\mathbb{P}(R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \geq \epsilon) \leq \delta$. The following result provides a lower-bound on the sample complexity of proper learning:

Theorem 4.4.3 (Thm. 11, [BHMZ20]). *Let $\epsilon \in (0, 1/8)$ and $\delta \in (0, 1/100)$. There exists a concept class with VC dimension d . for which we have $\mathcal{M}_{\text{prop}}^{\mathcal{H}}(\epsilon, \delta) \geq \frac{\tilde{c}}{\epsilon}(d \text{Log}_{\frac{1}{\epsilon}} + \text{Log}_{\frac{1}{\delta}})$ for a fixed numerical constant $\tilde{c} > 0$, where $\text{Log}(x) = \max\{1, \log(x)\}$ for $x \geq 0$.*

We now present the main result: for VC classes, we show that the existence of a learning algorithm with bounded CMI contradicts the lower bound on the sample complexity in Theorem 4.4.3. The proof can be found in Appendix C.3.

Theorem 4.4.4. *Statement 1 is false.*

Remark 4.4.5. Consider a modified Statement 1, seeking a proper learner with bounded eCMI instead. We can show that this modified statement is also false. \triangleleft

4.4.2 VC Classes with Finite Star Number

Theorem 4.4.4 states that it is not possible to find a proper learning algorithm with bounded CMI for *every* VC class. Note that this limitation does not imply a failure of the CMI framework for characterizing the expected excess risk of learning VC classes. Instead, the impossibility can be attributed to an inherent limitation of proper learning algorithms, since there exist VC classes such that no proper learning algorithm \mathcal{A}_n satisfies $\mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n)] = O(1/n)$ [BHMZ20]. In this section, we consider a family of VC classes for which we *can* show the existence of a learner with bounded CMI. We begin with some definitions.

Two sequences $((x_1, y_1), \dots, (x_n, y_n))$ and $((x'_1, y'_1), \dots, (x'_n, y'_n))$ are *neighbours* if $x_i = x'_i$ for all $i \in [n]$, and $y_i = y'_i$ for all but exactly one $i \in [n]$. Fix any concept class

$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. *Star number of \mathcal{H}* [HY15, Def. 2], denoted by \mathfrak{s} , is the largest integer n such that there exists a realizable $s \in (\mathcal{X} \times \mathcal{Y})^n$, and every neighbour of s is realizable by \mathcal{H} . If no such largest integer n exists, then $\mathfrak{s} = \infty$. Hanneke and Yang [HY15, Sec. 4.1] calculate the star number of some common concept classes. It is straightforward to see that $d \leq \mathfrak{s}$. For any $n \in \mathbb{N}$, and $s = ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, define a *version space* of s with respect to \mathcal{H} as $V_{\mathcal{H}}[s] = \{h \in \mathcal{H} : \hat{R}_s(h) = 0\}$, a set of classifiers that are consistent with s .

Star Number, Version Space, and CMI

Fix any concept class \mathcal{H} , and assume that, after observing a training sample S_n , we want to output the version space $V_{\mathcal{H}}[S_n]$, i.e., the set of all classifiers consistent with S_n . We are interested in the following question: for which concept classes does the version space carry little information about the training samples conditioned on the supersample? More precisely, for which classes is $I(V_{\mathcal{H}}[S_n]; U|Z) = O(1)$? Note that bounding the ‘‘CMI’’ of the version space provides a bound on the CMI of a broad class of algorithms that choose a particular ERM based solely on the version space, potentially under further constraints, such as privacy, fairness, etc.

In this section, we give a complete characterization of when $I(V_{\mathcal{H}}[S_n]; U|Z) = O(1)$, and show that it is possible if and only if \mathcal{H} has finite star number. In particular, given a class with infinite star number, we demonstrate that $I(V_{\mathcal{H}}[S_n]; U|Z) = \Omega(n)$. We begin with an upper bound, whose proof can be found in Appendix C.4.

Theorem 4.4.6. *Let $n \in \mathbb{N}$, \mathcal{H} be a concept class with star number \mathfrak{s} , and \mathcal{D} be a realizable distribution. Let Z , U , and S_n be as defined in the beginning of this section. Then for every $n \geq \mathfrak{s}$, we have $I(V_{\mathcal{H}}[S_n]; U|Z) \leq 2\mathfrak{s} \log 2$.*

We can use the data processing inequality and Theorem 4.4.6 to obtain the following:

Corollary 4.4.7. *Let \mathcal{H} be a concept class with the star number \mathfrak{s} . Consider any ERM algorithm \mathcal{A}_n for which the Markov chain $S_n - V_{\mathcal{H}}[S_n] - \mathcal{A}_n(S_n)$ holds; in other words, the output of the algorithm and the training set are conditionally independent given the version space. Then, for any such an algorithm, for every $n \geq \mathfrak{s}$, and every realizable distribution \mathcal{D} , we have $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 2\mathfrak{s} \log 2$.*

In Corollary 4.4.7, by assuming the Markov structure $S_n - V_{\mathcal{H}}[S_n] - \mathcal{A}_n(S_n)$ we restrict the information of the ERM algorithm $\mathcal{A}_n(S_n)$. One might try to extend our result in Corollary 4.4.7 such that it holds for *any* ERM without any constraints.

However, for the class of one-dimensional threshold over \mathbb{R} , whose star number is two, one can construct an ERM with maximal CMI [SZ20a, Sec. 4.3]. Therefore, the Markov chain assumption cannot be removed. The next theorem shows $\mathfrak{s} < \infty$ is a necessary condition, for otherwise, there exist learning scenarios under which we cannot output the version space, even with merely sublinear CMI.

Theorem 4.4.8. *For every $n \in \mathbb{N}$, $n \geq 2$ and for every concept class \mathcal{H} with star number \mathfrak{s} with $\mathfrak{s} \geq 2$ over input space \mathcal{X} , there exists a realizable data distribution \mathcal{D} on $\mathcal{X} \times \mathcal{Y}$ such that $I(V_{\mathcal{H}}[S_n]; U|Z) = \Omega(\min\{\mathfrak{s}, n\})$.*

Proof sketch: Let $\mathcal{X} = [n]$ and consider the concept class $\mathcal{H} = \{h_0, h_1, \dots, h_n : \mathcal{X} \rightarrow \mathcal{Y}\}$, where $h_0(x) = 0$ is the zero function and $h_t(x) = \mathbb{1}[x = t]$, for $t \in [n]$, are point functions. It is easy to see that this concept class has star number n on \mathcal{X} . Let \mathcal{D} correspond to the uniform distribution on \mathcal{X} and target function h_0 . Consider the bijection between \mathcal{H} and $\{0, 1, \dots, n\} \supseteq \mathcal{X}$. For every training sequence, the version space contains 0 and every point in \mathcal{X} not observed in $S_n = Z_U$. The key observation is that, in each column of Z , one point was *not* selected for training, and so each column contains zero or one points in the version space. Whenever there is one point, the value of U_i is revealed for that column. We show that the number of columns with this property is a lower bound on $I(V_{\mathcal{H}}[S_n]; U|Z)$. A coupon collector's argument yields a lower bound the number of such columns. The formal proof can be found in Appendix C.5. \square

An ERM whose CMI is logarithmic in star number

In the next theorem, we show that there exists an ERM for learning VC classes with a finite star number for which the CMI is upper bounded by a constant and its dependence on star number is logarithmic. The proof is provided in Appendix C.6.

Theorem 4.4.9. *Let \mathcal{H} be a concept class with VC dimension d and star number \mathfrak{s} . Then, there exists an ERM \mathcal{A}_n for learning \mathcal{H} such that for every $n \geq \mathfrak{s}$ and for every realizable distribution \mathcal{D} , we have $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) = O(d \log(\mathfrak{s}/d))$.*

Note that Theorem 4.4.9 shows the existence of a specific ERM with constant CMI, whereas in Corollary 4.4.7 we show a broad class of ERMs has bounded CMI.

4.5 Universality of eCMI and Improper Learning of VC Classes

The eCMI, introduced in Definition 6.7.3, is an appropriate information-theoretic notion for analyzing learning algorithms when there is no natural parameterization of the set of possible predictors, such as for improper or transductive algorithms. In this section, we show that eCMI is *universal* in the realizable setting. Then, we show that the CMI framework can be used to obtain a near-optimal bound on the expected excess risk of any algorithm with a leave-one-out error guarantee. As an application, we study CMI of the classical *one-inclusion graph prediction* algorithm, which was first proposed by Haussler, Littlestone, and M. K. Warmuth [HLW94] as an optimal improper learner for VC classes. The next theorem is the main result of this section, whose proof can be found in Appendix C.7.

Theorem 4.5.1. *Let $n \geq 2 \in \mathbb{N}$, let \mathcal{A}_n be a learning algorithm, and let \mathcal{D} be a distribution on \mathcal{Z} . Assume with probability one $\hat{R}_{S_n}(\mathcal{A}_n(S_n)) = 0$. Then,*

$$2/3R_{\mathcal{D}}(\mathcal{A}_n) \stackrel{(a)}{\leq} \text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))/n \stackrel{(b)}{\leq} H_b(R_{\mathcal{D}}(\mathcal{A}_n)) + R_{\mathcal{D}}(\mathcal{A}_n) \log(2), \quad (4.9)$$

where $H_b(\cdot)$ is the binary entropy function, and $R_{\mathcal{D}}(\mathcal{A}_n) = \mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n))]$.

The inequality (a) in Eq. (4.9) implies that, if $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))/n$ vanishes as n diverges, then $R_{\mathcal{D}}(\mathcal{A}_n)$ vanishes as well. The inequality (b) is more interesting: it implies that, if $R_{\mathcal{D}}(\mathcal{A}_n)$ vanishes as n diverges, then $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))/n$ also vanishes.

Assume that a consistent algorithm \mathcal{A} satisfies $R_{\mathcal{D}}(\mathcal{A}_n) = \theta/n$ for $\theta \in \mathbb{R} \geq 1$. Then, it is straightforward to see from Direction (b) in Eq. (4.9) that $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))/n = O(\theta \log(n))$. Also, for an algorithm with $R_{\mathcal{D}}(\mathcal{A}_n) = \theta \log(n)/n$ the upper bound in Eq. (4.9) is given by $O(\theta(\log(n))^2)$. This observation suggests that our upper bound for $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))/n$ in Eq. (4.9) provides a bound on the expected excess risk which is sub-optimal by a $\log(n)$ factor in some interesting cases.

Remark 4.5.2. Note that the result in Theorem 4.5.1 *does not* imply our results in former sections. In particular our results in Theorem 4.2.3, Theorem 4.3.4, Corollary 4.4.7, and (later in) Theorem 4.5.6 show that CMI framework provides *optimal* characterization of the expected excess risk in the considered scenarios. \triangleleft

The following corollary summarizes our result for the consistent algorithms with a leave-one-out error guarantee.

Corollary 4.5.3. *Let $n \in \mathbb{N}$ and $\theta \in \mathbb{R}_+$, such that $n \geq 2\theta$. Let \mathcal{A}_n be a consistent learning algorithm. Let \mathcal{D} be a distribution on \mathcal{Z} and assume that, with probability one over a sequence $S = (Z_1, \dots, Z_{n+1}) \sim \mathcal{D}^{n+1}$, we have $\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}^S[\ell(\mathcal{A}_n(S_{-i}), Z_i)] \leq \frac{\theta}{n+1}$, where the expectation is taken only over the randomness in \mathcal{A}_n . Then,*

$$\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) \leq \theta \log((n+1)/\theta) + 2\theta \log 2.$$

4.5.1 The One-Inclusion Graph Prediction Strategy

Haussler, Littlestone, and M. K. Warmuth [HLW94] proposed an improper learning rule for learning VC classes based on the one-inclusion graph [AHW87]. We provide a description of this algorithm in Appendix C.9. The deterministic version of this prediction rule satisfies the following property. Let \mathcal{H} be a concept class with VC dimension d . For every $n \in \mathbb{N}$, $h \in \mathcal{H}$, and $(x_1, \dots, x_{n+1}) \in \mathcal{X}^{n+1}$, let $S = ((x_1, h(x_1)), \dots, (x_{n+1}, h(x_{n+1})))$. Then $\frac{1}{n+1} \sum_{i=1}^{n+1} \ell(\mathcal{A}_n(S_{-i}), (x_i, h(x_i))) \leq \frac{d}{n+1}$. A direct application of Corollary 4.5.3 gives the following results.

Corollary 4.5.4. *Let \mathcal{A}_n denote the deterministic one-inclusion graph for learning class \mathcal{H} with VC dimension d . Then, for every realizable distribution \mathcal{D} and $n \geq 2d$, we have $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) \leq d \log((n+1)/d) + 2d \log 2$.*

Remark 4.5.5. In Theorem 4.2.3 we provide a bound on eCMI of any proper ERM. However, for improper learners, we can construct a consistent algorithm with maximal eCMI. For instance, consider $\mathcal{X} = [0, 1]$, $\mathcal{D}_X = \text{Unif}([0, 1])$, the concept class of threshold with target function $h^*(x) = \mathbb{1}[x \geq 1/2]$. Consider a learning algorithm that gives the correct predictions on the points that are in the training set, and for a point that is not in the training set it always predicts one. One can show that eCMI of this consistent algorithm is $\Omega(n)$. \triangleleft

Haussler, Littlestone, and M. K. Warmuth [HLW94] showed that the one-inclusion graph algorithm achieves $\mathbb{E}R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \leq d/n$ for learning a class \mathcal{H} with VC dimension d . Corollary 4.5.4 implies that $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) = O(d \log(n))$ for every *deterministic* one-inclusion graph prediction rule. Combining this result with Eq. (4.3) provides a bound on the excess risk which is suboptimal by a $\log n$ factor. In the next theorem, we show that, in at least one interesting special case, it is possible to remove the logarithmic factor from eCMI by exploiting a *randomized* one-inclusion graph prediction algorithm.

Theorem 4.5.6. *Let \mathcal{H} denote the class of singletons (point functions) on $\mathcal{X} = \mathbb{R}$. There exists a randomized one-inclusion graph prediction rule \mathcal{A}_n for learning class \mathcal{H} such that for every realizable distribution \mathcal{D} and $n \geq 2$, we have $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) = O(1)$.*

Chapter 5

Understanding Generalization via Leave-One-Out Conditional Mutual Information

5.1 Introduction

In this chapter, we introduce a new information-theoretic measure of dependence between the output of a learning algorithm and its input, based on a leave-one-out analogue of eCMI:

Definition 5.1.1. Let \mathcal{A} be a learning algorithm. Let $n \in \mathbb{N}$, let $\tilde{Z} = (\tilde{Z}_i)_{i \in [n+1]}$ be an $n + 1$ -array of i.i.d. random elements in the dataspace \mathcal{Z} with common distribution \mathcal{D} . Let U be a random variables distributed uniformly on $[n + 1]$, independent from \tilde{Z} . For $u \in [n + 1]$, let \tilde{Z}_{-u} denote $(\tilde{Z}_j)_{j \in [n+1], j \neq u}$, i.e., the supersample with the u 'th element removed. Define $S_n = \tilde{Z}_{-U}$. Let $L \in \mathbb{R}_+^{n+1}$ be the array with entries $L_i = \ell(\mathcal{A}_n(S_n), \tilde{Z}_i)$ for $i \in [n + 1]$. The *leave-one-out (evaluated) conditional mutual information of \mathcal{A}_n with respect to \mathcal{D}* is

$$\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \triangleq I(L; U | \tilde{Z}).$$

Our notion is inspired by the *leave-one-out error* estimator of the generalization error [MRT18], a widely used surrogate for the expected generalization error. Intuitively, $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ measures how well one can “identify” which point from the supersample is being held-out, given the supersample and the losses incurred on each of its elements. In Section 6.4, we show that bounded $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ implies generalization in both interpolating and agnostic learning algorithms. Our generalization

bound for the interpolating case enjoys the property that it is never greater than one.

In Section 5.3, in the context of 0–1 loss and interpolating learning algorithms, we establish a general connection between LOO^eCMI and the classical leave-one-out error estimator of the risk. Using this result, we show that $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ precisely characterizes the risk of interpolating learning algorithms in many common situations. Specifically, we show that, for every data distribution and interpolating learner, the $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ framework yields a risk bound that vanishes *if and only if* the risk also vanishes as the number training samples diverges. Also, the $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ framework is the first information-theoretic framework that can be shown to determine the risk of any consistent learning algorithms whose risk either converge to a non-zero value or converges to zero polynomially with the number of samples.

In Section 5.4, as an application of our general connection with leave-one-out error analysis, we characterize the LOO^eCMI of the one-inclusion graph algorithm, which was introduced by Haussler, Littlestone, and M. K. Warmuth [HLW94] for learning Vapnik–Chervonenkis (VC) classes. Using our framework, we obtain the *optimal* risk bound for every VC class in the realizable setting. In doing so, we answer the open problem stated in [SZ20b] of characterizing the expected excess risk of learning VC classes using an information-theoretic framework.

In Section 5.5, we consider several additional measures of information based on the supersample structure introduced in Definition 5.1.1, show that they all control generalization, and discuss their inter-relationships. In particular, we present the chain of inequalities

$$I(L; U) \leq I(L; U|\tilde{Z}) \leq I(\hat{Y}; U|\tilde{Z}) \leq I(\mathcal{A}_n(S_n); U|\tilde{Z}) \leq I(\mathcal{A}_n(S_n); S_n), \quad (5.1)$$

where $I(\mathcal{A}_n(S_n); U|\tilde{Z})$ is the non-evaluated analogue of our notion and \hat{Y} is the length- $n + 1$ list of labels predicted by $\mathcal{A}_n(S_n)$ on the inputs in the supersample \tilde{Z} . With the exception of the first quantity, $I(L; U)$, all of these (or close analogues) have been studied in the literature. In the special case of binary classification by an interpolating classifier, $I(L; U)$ is precisely the risk, yielding a simple argument for why the other notions also bound risk (equivalently, generalization error) in this setting. Based on results presented herein and elsewhere, we discuss gaps between these various notions in Eq. (5.1) and the roles they can play in understanding generalization. In particular, we show that LOO^eCMI is the weakest measure in this chain that can characterize the risk for every interpolating algorithm under 0–1 loss.

5.1.1 Related Work

For supervised learning algorithms, Steinke and Zakyntinou [SZ20a, Sec. 6] and Harutyunyan, Raginsky, Ver Steeg, and Galstyan [HRVG21] define information-theoretic measures of dependence based on the losses and predictions, respectively, of a learning algorithm rather than the learned classifier. The results in both of these papers are based on the CMI framework, i.e., a supersample with $2n$ samples. Neither of these papers recover optimal bounds for learning VC classes in the realizable setting. Hafez-Kolahi, Golgooni, Kasaei, and Soleymani [HGKS20] combine chaining with CMI to study generalization of deterministic learning algorithms. For ERM on classes of finite VC dimension d in the *agnostic* case, i.e., $\inf_{h \in \mathcal{H}} R_{\mathcal{D}}(h) \neq 0$, Hafez-Kolahi, Golgooni, Kasaei, and Soleymani use chaining CMI to obtain bounds on the expected generalization error that achieve the optimal rate $O(\sqrt{d/n})$. See also [ATR21; CSDD22; ZTL21; HMK22; IEG19; JHW17; AAV18; ZTL22; RBTS20].

5.1.2 Notation

Let P, Q be probability measures. For a P -integrable function f , let $P[f] = \int f dP$. When Q is absolutely continuous with respect to P , denoted $Q \ll P$, write $\frac{dQ}{dP}$ for (an arbitrary version of) the Radon–Nikodym derivative (or density) of Q with respect to P . The *KL divergence* (or *relative entropy*) of Q with respect to P , denoted $\text{KL}(Q \parallel P)$, is defined as $Q[\log \frac{dQ}{dP}]$ when $Q \ll P$ and infinity otherwise.

For a random element X in some measurable space \mathcal{X} , let $\mathbb{P}[X]$ denote its distribution, which lives in the space $\mathcal{M}_1(\mathcal{X})$ of all probability measures on \mathcal{X} . Given another random element, say Y in \mathcal{T} , let $\mathbb{P}^Y[X]$ denote the conditional distribution of X given Y . If X and Y are independent, denoted by $X \perp\!\!\!\perp Y$, we have $\mathbb{P}^Y[X] = \mathbb{P}[X]$ a.s.

The *mutual information between X and Y* is $I(X; Y) = \text{KL}(\mathbb{P}[(X, Y)] \parallel \mathbb{P}[X] \otimes \mathbb{P}[Y])$, where \otimes forms the product measure. Writing $\mathbb{P}^Z[(X, Y)]$ for the conditional distribution of the pair (X, Y) given a random element Z , the *disintegrated mutual information between X and Y given Z* , is $I^Z(X; Y) = \text{KL}(\mathbb{P}^Z[(X, Y)] \parallel \mathbb{P}^Z[X] \otimes \mathbb{P}^Z[Y])$, and the conditional mutual information is $I(X; Y|Z) = \mathbb{E}[I^Z(X; Y)]$. Similarly, we can define *disintegrated entropy of X given Y* denoted by $H^Y(X)$, and its expectation gives the *conditional entropy of X given Y* , i.e., $H(X|Y) = \mathbb{E}[H^Y(X)]$.

5.2 Generalization Bounds

In this section, we show that bounded LOO^eCMI implies generalization. First, we provide a generalization bound for interpolating learning algorithms:

Theorem 5.2.1. *Let $n \in \mathbb{N}$, assume loss is bounded in $[0, 1]$, and let \mathcal{A}_n be an interpolating learning algorithm. Then,*

$$R_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{\log(n+1)}.$$

Remark 5.2.2. By the definition of mutual information, $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq H(U|\tilde{Z}) \leq \log(n+1)$, since $U \perp\!\!\!\perp \tilde{Z}$ and U is distributed uniformly on $[n+1]$. Therefore, the bound above is never greater than one. \triangleleft

Proof of Theorem 5.2.1. Let \mathcal{D} , \tilde{Z} , U , and L be defined as in Definition 5.1.1. Introduce \tilde{U} such that $\tilde{U} \stackrel{d}{=} U$ and $\tilde{U} \perp\!\!\!\perp (\tilde{Z}, L)$. By the Donsker–Varadhan variational formula [BLM13, Prop. 4.15], for all bounded measurable functions h and $\lambda \in \mathbb{R}$,

$$\begin{aligned} I(\tilde{Z}, L; U) &= \text{KL}(\mathbb{P}(\tilde{Z}, U, L) \parallel \mathbb{P}(\tilde{Z}, L) \otimes \mathbb{P}(\tilde{U})) \\ &\geq \mathbb{P}(\tilde{Z}, U, L)(\lambda h) - \log [(\mathbb{P}(\tilde{Z}, L) \otimes \mathbb{P}(\tilde{U}))(\exp(\lambda h))]. \end{aligned} \quad (5.2)$$

Let $\alpha \in \mathbb{R}_+$ be a constant. Consider the function $h_\alpha : \mathcal{Z}^{n+1} \times [n+1] \times [0, 1]^{n+1} \rightarrow \mathbb{R}$ given by $h_\alpha(\tilde{z}, u, l) = l_u - \alpha \sum_{i \in [n+1], i \neq u} l_i$. Then

$$\begin{aligned} \mathbb{P}(\tilde{Z}, U, L)(h_\alpha) &= \mathbb{E}[\mathbb{E}^U h_\alpha(U, L, \tilde{Z})] \\ &= \mathbb{E}[\mathbb{E}^U [\ell(\mathcal{A}_n(\tilde{Z}_{-U}), \tilde{Z}_U) - \alpha \sum_{i \in [n], i \neq U} \ell(\mathcal{A}_n(\tilde{Z}_{-U}), \tilde{Z}_i)]] \\ &\stackrel{(a)}{=} \mathbb{E}[\mathbb{E}^U [\ell(\mathcal{A}_n(\tilde{Z}_{-U}), \tilde{Z}_U)]] \\ &= \mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n))]. \end{aligned} \quad (5.3)$$

Here, (a) follows from the fact that \mathcal{A}_n is an interpolating algorithm, hence, for all $i \neq U$, $\ell(\mathcal{A}_n(\tilde{Z}_{-U}), \tilde{Z}_i) = 0$ a.s. Thus, by Eq. (5.2), Eq. (5.3), and $I(L, \tilde{Z}; U) = \text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$, we obtain

$$\begin{aligned} R_{\mathcal{D}}(\mathcal{A}_n) &\leq \frac{\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{\lambda} \\ &\quad + \frac{\log [\mathbb{P}(\tilde{Z}, L) \otimes \mathbb{P}(\tilde{U})(\exp(\lambda h_\alpha))]}{\lambda}. \end{aligned} \quad (5.4)$$

To simplify the notation, let $\ell(i, j) = \ell(\mathcal{A}_n(\tilde{Z}_{-i}, \tilde{Z}_j))$ for i and j in $[n + 1]$. We have

$$\begin{aligned}
& \mathbb{P}(\tilde{Z}, L) \otimes \mathbb{P}(\tilde{U})(\exp(\lambda h_\alpha)) \\
& \stackrel{(a)}{=} \mathbb{E} \left[\exp \left[\lambda(\ell(U, \tilde{U}) - \alpha \sum_{i \in [n+1], i \neq \tilde{U}} \ell(U, i)) \right] \right] \\
& = \mathbb{E} \left[\mathbb{1}[U = \tilde{U}] \exp \left[\lambda(\ell(U, \tilde{U}) - \alpha \sum_{i \in [n+1], i \neq \tilde{U}} \ell(U, i)) \right] \right. \\
& \quad \left. + \mathbb{E} \left[\mathbb{1}[U \neq \tilde{U}] \exp \left[\lambda(\ell(U, \tilde{U}) - \alpha \sum_{i \in [n+1], i \neq \tilde{U}} \ell(U, i)) \right] \right] \right] \tag{5.5} \\
& \stackrel{(b)}{=} \mathbb{E} \left[\mathbb{1}[U = \tilde{U}] \exp(\lambda \ell(U, U)) + \mathbb{1}[U \neq \tilde{U}] \exp(-\lambda \alpha \ell(U, U)) \right] \\
& \stackrel{(c)}{=} \mathbb{E} \left[\frac{1}{n+1} \exp(\lambda \ell(U, U)) + \frac{n}{n+1} \exp(-\lambda \alpha \ell(U, U)) \right].
\end{aligned}$$

The equality (a) follows from the chain rule. In particular, the expectation over \tilde{Z} and L can be written as $\mathbb{E} \mathbb{E}^{\tilde{Z}, U}$. Step (b) follows from the fact that for all $i \in [n + 1]$ and $i \neq U$, $\ell(U, i) = 0$ a.s., since the algorithm is interpolating. In (c) we take the expectation with respect to \tilde{U} and use the fact that \tilde{U} is independent of \tilde{Z} , U , and the internal randomness of \mathcal{A} . Let $\lambda = \log(n + 1)$. Consider the function $f : [0, 1] \rightarrow \mathbb{R}$, where $f(x) = \frac{1}{n+1} \exp(\lambda x) + \frac{n}{n+1} \exp(-\lambda \alpha x)$. Note that f is the sum of two convex functions, defined on a bounded domain. It achieves its maximum over the endpoint, i.e., $x \in \{0, 1\}$. Considering this observation, we can further upper bound Eq. (5.5) by considering the following two cases: If $\ell(U, U) = 1$, then $\frac{\exp(\lambda \ell(U, U)) + n \exp(-\lambda \alpha \ell(U, U))}{n+1} = 1 + \frac{n}{n+1} \exp(-\alpha \log(n + 1))$. Otherwise, in the case that $\ell(U, U) = 0$, we have

$$\frac{\exp(\lambda \ell(U, U)) + n \exp(-\lambda \alpha \ell(U, U))}{n+1} = 1.$$

Therefore, we conclude

$$\mathbb{P}(\tilde{Z}, L) \otimes \mathbb{P}(\tilde{U})(\exp(\lambda h_\alpha)) \leq 1 + \frac{n}{n+1} \exp(-\alpha \log(n + 1)). \tag{5.6}$$

By Eq. (5.4) and Eq. (5.6)

$$R_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{I(L; U | \tilde{Z}) + \log \left(1 + \frac{n \exp(-\alpha \log(n+1))}{n+1} \right)}{\log(n+1)}. \tag{5.7}$$

Finally, letting $\alpha \rightarrow +\infty$ in Eq. (5.7) concludes the proof. \square

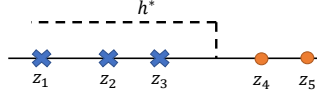


Figure 5.1: A counter example.

Remark 5.2.3. The special case of zero–one loss can be proven in a more direct way. See Section 5.5. \triangleleft

Remark 5.2.4. In this remark, we want to show that in general, conditional on the event $L = (0, \dots, 0)$, it does not hold that $\mathbb{E}\left[H^{L, \tilde{Z}}(U) | L = (0, \dots, 0)\right] = \log(n + 1)$.

Consider the class of thresholds in one dimension. In Fig. 5.1, we consider learning from $n = 4$ data, when the supersample is in a neighborhood of a sequence (z_1, \dots, z_5) . Given the training data S_n , let $x^* = \max\{x | (x, 1) \in S_n\}$ if S_n contains at least one point with label 1, otherwise let $x^* = -\infty$. Consider the learning algorithm $\mathcal{A}_n(S_n) = \hat{h}$ where $\hat{h}(x) = \mathbb{1}[x \leq x^*]$. On the event $L = (0, \dots, 0)$ and \tilde{Z} is a small perturbation of the points (z_1, \dots, z_5) , we have $\mathbb{P}^{L, \tilde{Z}}[U = 3] = 0$. The reason is, had it been the case that $U = 3$, then the learning algorithm would have erred on its prediction for (the point corresponding to) z_3 . Therefore, $\mathbb{E}\left[H^{L, \tilde{Z}}(U) | L = (0, \dots, 0)\right] \neq \log(n + 1)$, in general. \triangleleft

For an arbitrary learning algorithm, LOO^eCMI still controls generalization. The proof of the following theorem is deferred to Section 5.5.

Theorem 5.2.5. *Let $n \in \mathbb{N}$. Assuming only that loss is bounded in $[0, 1]$,*

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \sqrt{2\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}.$$

5.3 A Connection with leave-one-out error

In this section, we describe a connection between LOO^eCMI and the *leave-one-out error*, a well-studied statistical estimator of risk [Cov69; Sto76]. Using the supersample \tilde{Z} , the random variable

$$\hat{R}_{\text{loo}} = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}^{\tilde{Z}} \left[\ell(\mathcal{A}_n(\tilde{Z}_{-i}), Z_i) \right]$$

is a leave-one-out error estimate for the risk of \mathcal{A}_n , where we have averaged out the internal randomness in \mathcal{A}_n . (Note that, for deterministic learning algorithms, this averaging has no effect.) In order to connect this quantity to LOO^eCMI , we first note

that we can bound $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ in terms of the disintegrated entropy:

$$\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) = \mathbb{E} \left[\mathbb{H}^{\tilde{Z}}(L) - \mathbb{H}^{\tilde{Z},U}(L) \right] \leq \mathbb{E} \left[\mathbb{H}^{\tilde{Z}}(L) \right], \quad (5.8)$$

where the second inequality is an equality for deterministic learning algorithms. Let $\mathbb{H}_b(\cdot)$ denote the binary entropy function.

Theorem 5.3.1. *Let \mathcal{A}_n be an interpolating learning algorithm for some data distribution \mathcal{D} . Assume loss lies in $\{0, 1\}$. Then, almost surely,*

$$\mathbb{H}^{\tilde{Z}}(L) \leq \mathbb{H}_b(\hat{R}_{\text{loo}}) + \hat{R}_{\text{loo}} \log(n+1). \quad (5.9)$$

Proof. Let $0_{(0)} = (0, \dots, 0) \in \{0, 1\}^{n+1}$ and, for $i \in \{1, \dots, n+1\}$, let $0_{(i)} \in \{0, 1\}^{n+1}$ be equivalent to $0_{(0)}$ but for a 1 at index i . Due to interpolation, the support of L is $\{0_{(i)} | i \in \{0, \dots, n+1\}\}$. For each $i, j \in \{1, \dots, n+1\}$, let

$$\kappa_{i,j} = \mathbb{P}^{\tilde{Z},U=j}[L = 0_{(i)}] = \mathbb{P}^{\tilde{Z}}[\ell(\mathcal{A}_n(\tilde{Z}_{-j}), \tilde{Z}_i) = 1].$$

Since \mathcal{A}_n is a consistent algorithm $\kappa_{i,j} = 0$ a.s. for $i \neq j$. Also, $\mathbb{E}[\kappa_{i,i}] = R_{\mathcal{D}}(\mathcal{A}_n)$. For $i \in \{1, \dots, n+1\}$,

$$\begin{aligned} \mathbb{P}^{\tilde{Z}}[L = 0_{(i)}] &= \mathbb{E}^{\tilde{Z}}[\mathbb{P}^{\tilde{Z},U}[L = 0_{(i)}]] \\ &= \frac{1}{n+1} \sum_{j=1}^{n+1} \kappa_{i,j} \\ &= \frac{\kappa_{i,i}}{n+1}, \end{aligned}$$

where the last line follows since $\kappa_{i,j} = 0$ for $i \neq j$. Note that the leave-one-out-error satisfies $\hat{R}_{\text{loo}} = (n+1)^{-1} \sum_{i=1}^{n+1} \kappa_{i,i}$. Then, by the definition of the entropy,

$$\begin{aligned} \mathbb{H}^{\tilde{Z}}(L) &= - \sum_{i=0}^{n+1} \mathbb{P}^{\tilde{Z}}[L = 0_{(i)}] \log \mathbb{P}^{\tilde{Z}}[L = 0_{(i)}] \\ &= -(1 - \hat{R}_{\text{loo}}) \log(1 - \hat{R}_{\text{loo}}) - \sum_{i=1}^{n+1} \frac{\kappa_{i,i}}{n+1} \log\left(\frac{\kappa_{i,i}}{n+1}\right). \end{aligned} \quad (5.10)$$

We now invoke the log-sum inequality for non-negative sequences $\{a_i\}_{i \in [n]}$ and $\{b_i\}_{i \in [n]}$, wherein $\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq (\sum_{i \in [n]} a_i) \log \frac{\sum_{i \in [n]} a_i}{\sum_{i \in [n]} b_i}$. Using this inequality, we

obtain

$$\sum_{i=1}^{n+1} \frac{\kappa_{i,i}}{n+1} \log \left(\frac{\kappa_{i,i}}{n+1} \right) \geq \hat{R}_{\text{loo}} \log \left(\frac{\hat{R}_{\text{loo}}}{n+1} \right). \quad (5.11)$$

Therefore, by Eqs. (5.10) and (5.11),

$$\begin{aligned} \mathbb{H}^{\tilde{Z}}(L) &\leq -(1 - \hat{R}_{\text{loo}}) \log(1 - \hat{R}_{\text{loo}}) \\ &\quad - \hat{R}_{\text{loo}} \log \hat{R}_{\text{loo}} + \hat{R}_{\text{loo}} \log(n+1) \\ &= \mathbb{H}_b(\hat{R}_{\text{loo}}) + \hat{R}_{\text{loo}} \log(n+1), \end{aligned}$$

as was to be shown. \square

As a corollary, we provide an explicit bound on LOO^eCMI.

Corollary 5.3.2. *Let \mathcal{A}_n be a consistent learning algorithm for zero-one valued loss, let \mathcal{D} be a distribution on \mathcal{Z} , and assume that, with probability one, $\hat{R}_{\text{loo}} \leq \frac{\theta}{n+1}$, where θ is some \tilde{Z} -measurable random variable in \mathbb{R}_+ . Then, almost surely,*

$$I^{\tilde{Z}}(L; U) \leq \begin{cases} 1 + \frac{\theta \log(n+1)}{n+1}, & \text{if } \frac{\theta}{n+1} \geq \frac{1}{2}, \\ \frac{2\theta \log(n+1)}{n+1} + \frac{\theta + \exp(-1)}{n+1}, & \text{otherwise.} \end{cases}$$

Proof. For the case $2\theta \geq n+1$, upper-bounding the $\mathbb{H}_b(\hat{R}_{\text{loo}})$ by one, we obtain the result. For the case $2\theta < n+1$, note that $\mathbb{H}_b(\hat{R}_{\text{loo}}) \leq \mathbb{H}_b(\frac{\theta}{n+1})$. Using the well-known inequality $\mathbb{H}_b(x) \leq -x \log(x) + x$, we obtain

$$\begin{aligned} I^{\tilde{Z}}(L; U) &\leq \frac{\theta}{n+1} \log \frac{n+1}{\theta} + \frac{\theta}{n+1} + \frac{\theta}{n+1} \log(n+1) \\ &= \frac{\theta}{n+1} (\log(n+1) - \log(\theta)) + \frac{\theta}{n+1} + \frac{\theta \log(n+1)}{n+1} \\ &= \frac{2\theta \log(n+1)}{n+1} - \frac{\theta \log(\theta)}{n+1} + \frac{\theta}{n+1} \\ &\leq \frac{2\theta \log(n+1)}{n+1} + \frac{\exp(-1)}{n+1} + \frac{\theta}{n+1}, \end{aligned}$$

where the last line follows from $\max_{x>0} -x \log(x) = \exp(-1)$. \square

Corollary 5.3.2 can be used to obtain risk bound for a variety of consistent learning algorithms. For example, the Support Vector Machine (SVM) algorithm has a leave-one-out error guarantee in the sense of Corollary 5.3.2 for realizable distributions, where θ is given by the number $N_{\text{SV}}(\tilde{Z})$ of support vectors in the supersample \tilde{Z}

[MRT18]. Therefore, assuming \mathcal{D} is a realizable distribution with respect to the class of half-spaces in \mathbb{R}^d , the LOO^eCMI of the SVM satisfies

$$\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{\mathbb{E}[N_{\text{SV}}(\tilde{Z})]}{n+1} (2\log(n+1) + 1).$$

Combining this result with Theorem 5.2.1 yields a bound on the risk of SVM that is optimal up to a constant factor [LL20].

5.3.1 Universality of LOO^eCMI

In this section, we demonstrate that leave-one-out CMI captures the asymptotics of risk for consistent learners. More precisely, for every data distribution \mathcal{D} and consistent learner \mathcal{A} , the quantity $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)/\log(n+1)$ vanishes as the number training samples n diverges if and only if the risk also vanishes. Also, for a broad class of learning algorithms for which the population risk converges to zero polynomially in the size of the training set, $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)/\log(n+1)$ vanishes at the same rate as $R_{\mathcal{D}}(\mathcal{A}_n)$.

Theorem 5.3.3. *Let \mathcal{A}_n be an interpolating learning algorithm for some data distribution \mathcal{D} . Assume loss lies in $\{0, 1\}$. Then,*

$$\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq H_b(R_{\mathcal{D}}(\mathcal{A}_n)) + R_{\mathcal{D}}(\mathcal{A}_n) \log(n+1) \quad (5.12)$$

and

$$R_{\mathcal{D}}(\mathcal{A}_n) \log(n+1) \leq \text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n). \quad (5.13)$$

Proof. Since the binary entropy function is concave, it follows from Jensen's inequality, Theorem 5.3.1, and the identity $\mathbb{E}[\hat{R}_{\text{loo}}] = R_{\mathcal{D}}(\mathcal{A}_n)$ that

$$\begin{aligned} H(L|\tilde{Z}) &= \mathbb{E}[H^{\tilde{Z}}(L)] \\ &\leq \mathbb{E}[H_b(\hat{R}_{\text{loo}}) + \hat{R}_{\text{loo}} \log(n+1)] \\ &\leq H_b(R_{\mathcal{D}}(\mathcal{A}_n)) + R_{\mathcal{D}}(\mathcal{A}_n) \log(n+1), \end{aligned}$$

which was to be shown. Finally, the lower bound (Eq. (5.13)) is Theorem 5.2.1. \square

Remark 5.3.4. Are these bounds tight for consistent learners in the large n limit? First consider the case that the risk of \mathcal{A}_n does not converge to zero as n diverges,

i.e., $R_{\mathcal{D}}(\mathcal{A}_n) = \Theta(1)$. Then $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)/\log(n+1) = \Theta(R_{\mathcal{D}}(\mathcal{A}_n))$. For consistent learning algorithms such that $R_{\mathcal{D}}(\mathcal{A}_n) = c \frac{\log(n)^\alpha}{n^\beta}$ where c , α and β are some non-negative constants, we claim that $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)/\log(n+1) = \Theta(R_{\mathcal{D}}(\mathcal{A}_n))$. This claim follows by bounding Eq. (5.12) using the well-known inequality $H_b(p) \leq -p \log(p) + p$ for $p \in [0, 1]$. \triangleleft

Remark 5.3.5. Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \{0, 1\}$. The class of (right-continuous) one-dimensional thresholds over \mathcal{X} is $\mathcal{H} = \{h_\theta | \theta \in \mathcal{X}\}$ where $h_\theta(x) = \mathbb{1}[x > \theta]$. Consider a continuous realizable distribution D for \mathcal{H} with positive margin, i.e., the data labelled 0 and 1 are separated by an interval. For any proper algorithm, we may write $\mathcal{A}(S_n) = h_{\hat{\theta}}$ for some random variable $\hat{\theta}$ in \mathcal{X} . Using the margin assumption, we can design \mathcal{A} so that (1) $h_{\hat{\theta}}$ achieves zero training error yet (2) the representation of $\hat{\theta}$ encodes the whole training set, in the sense that, having $\hat{\theta}$ and \tilde{Z} , we can decode U perfectly. For this algorithm, we then have $I(\mathcal{A}(S_n); U | \tilde{Z}) = \log(n+1)$. On the other hand, for this class, [Han16] showed that any algorithm with zero training error achieves $R_{\mathcal{D}}(\mathcal{A}_n) = O(1/n)$. Therefore, our result in Theorem 5.3.3 shows that $I(L; U | \tilde{Z})/\log(n+1) = O(1/n)$. This example separates $I(L; U | \tilde{Z})$ and $I(\mathcal{A}(S_n); U | \tilde{Z})$, and served as motivation for Definition 5.1.1. As we discuss elsewhere, if sufficiently tight control on generalization can be obtained via a formally looser bound, doing so yields a stronger explanation. \triangleleft

5.4 An Optimal Bound for Learning VC classes

We start this section with some standard definitions [SB14]. Consider binary classification, i.e., $\mathcal{Y} = \{0, 1\}$, with zero-one loss. A sequence $((x_1, y_1), \dots, (x_n, y_n))$ is said to be *realizable by \mathcal{H}* , if for some $h \in \mathcal{H}$, $h(x_i) = y_i$ for all $i \in [n] = \{1, \dots, n\}$. Note that, if \mathcal{D} is realizable by \mathcal{H} , then, for all $n \in \mathbb{N}$, the sequence $S_n \sim \mathcal{D}^n$ is a.s. realizable by \mathcal{H} . We say \mathcal{H} *shatters* $(x_1, \dots, x_m) \in \mathcal{X}^m$ if for all $(y_1, \dots, y_m) \in \{0, 1\}^m$, there exists $h \in \mathcal{H}$, such that, for all $i \in [m]$, we have $h(x_i) = y_i$. The *VC dimension* of \mathcal{H} , denoted d , is the supremum of integers $m \geq 0$ for which there exists $(x_1, \dots, x_m) \in \mathcal{X}^m$ shattered by \mathcal{H} . In particular, if there is no largest such integer, the VC dimension is infinite, i.e., $d = \infty$.

In this section, we study the LOO^eCMI of the classical one-inclusion graph algorithm, which was first proposed by Haussler, Littlestone, and M. K. Warmuth [HLW94] as a general-purpose transductive learner for VC classes in the realizable setting. Here, we provide a brief description of this algorithm. Let \mathcal{H} be a class with a bounded VC dimension. Assume a realizable sequence of n la-

beled samples $S_n = ((x_1, y_1), \dots, (x_n, y_n)) \in \mathcal{Z}^n$ and a test point x_{n+1} are given to the learner, and the learner is tasked to predict the label of x_{n+1} . Let V be the set of all possible labelings of (x_1, \dots, x_{n+1}) by the classifiers in \mathcal{H} , i.e., $V = \{(v_1, \dots, v_{n+1}) \in \{0, 1\}^{n+1} \mid \exists h \in \mathcal{H}, \forall i \in [n+1], h(x_i) = v_i\}$. The *one-inclusion graph* [HLW94] is the graph with vertex set V such that vertices $\vec{v}, \vec{w} \in V$ are connected by an edge if and only if the Hamming distance between \vec{v} and \vec{w} is one. A *probability assignment* is a function $P : V \times V \rightarrow [0, 1]$ such that (i) $P(\vec{v}, \vec{w}) > 0$ only if \vec{v} and \vec{w} are adjacent (in particular, $P(\vec{v}, \vec{v}) = 0$ for all $\vec{v} \in V$) and (ii) given two adjacent vertices \vec{v} and \vec{w} , we have $P(\vec{v}, \vec{w}) + P(\vec{w}, \vec{v}) = 1$. We assume that P is chosen based only on (x_1, \dots, x_{n+1}) , i.e., independently of the labels.

We say a vertex $\vec{v} = (v_1, \dots, v_{n+1}) \in V$ is consistent with the labels in S_n if $(v_1, \dots, v_n) = (y_1, \dots, y_n)$. Since the labels of (x_1, \dots, x_n) are known, Haussler, Littlestone, and M. K. Warmuth [HLW94] observed that at most *two* vertices in the one-inclusion graph are consistent with the labels in S_n . In the case that only vertex $\vec{v} = (v_1, \dots, v_{n+1}) \in V$ is consistent with S_n , the label of x_{n+1} is predicted as v_{n+1} . In the case that two vertices $\vec{v} = (v_1, \dots, v_{n+1})$ and $\vec{w} = (w_1, \dots, w_{n+1})$ are consistent with S_n , they differ only on the $(n+1)$ -th position and the algorithm uses the probability assignment P to predict that the label for x_{n+1} agrees with v_{n+1} with probability $P(\vec{v}, \vec{w})$ and agrees with w_{n+1} otherwise.

Consider a realizable distribution \mathcal{D} and let $h^* \in \mathcal{H}$ denote a function that determines the labels. Haussler, Littlestone, and M. K. Warmuth [HLW94] prove that the leave-one-out error of the one-inclusion graph algorithm can be expressed in terms of the probability assignment P as

$$\hat{R}_{\text{loo}} = \frac{\sum_{\vec{v} \in V} P(\vec{v}^*, \vec{v})}{(n+1)},$$

where $\vec{v}^* = (h^*(X_1), \dots, h^*(X_{n+1})) \in V$ is the vertex corresponding to h^* in the one-inclusion graph of (X_1, \dots, X_{n+1}) . Moreover, Haussler, Littlestone, and M. K. Warmuth [HLW94] prove that there exists a probability assignment P such that $\sum_{\vec{v} \in V} P(\vec{w}, \vec{v}) \leq d$ uniformly over all $\vec{w} \in V$. By combining this result and Theorem 5.3.1, we obtain the main result of this section.

Theorem 5.4.1. *Let \mathcal{A} denote the one-inclusion graph algorithm. Then, for every VC class \mathcal{H} with dimension d , every data distribution \mathcal{D} realizable by \mathcal{H} , and $n \geq d$, there exists a probability assignment for the one-inclusion graph algorithm such that $\text{LOO}^e\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{d}{n+1}(2 \log(n+1) + 1)$. Combining this results with the generalization*

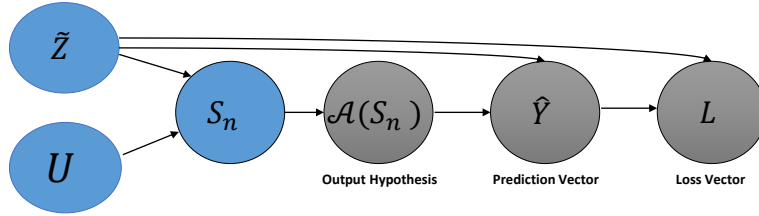


Figure 5.2: Conditional independence relationships encoded as a graphical model.

bound in Theorem 5.2.1 yields expected risk $\frac{2d}{n+1}(1 + o(1))$ which is optimal up to a constant factor [LLS01].

Using our general reduction (Theorem 5.3.1), we have characterized the $\text{LOO}^{\text{e}}\text{CMI}$ of the one-inclusion graph algorithm and, as a consequence, shown that the $\text{LOO}^{\text{e}}\text{CMI}$ framework provides an *optimal* bound for learning VC classes. It is worth noting that the IOMI framework of Russo and J. Zou [RZ16] and Xu and Raginsky [XR17] is provably *unable* to characterize the learnability of VC classes [LM20], and, at present, the best known bound within the CMI framework [SZ20a] is *suboptimal* by a $\log(n)$ factor [SZ20b].

5.5 A Hierarchy of Measures of Information

We have introduced leave-one-out CMI and shown that it can be used to bound the expected generalization error of learning algorithms. In this section, we aim to relate leave-one-out CMI to other measures of mutual information, some of which have already been shown to control generalization.

To begin, we place leave-one-out CMI in a chain of inequalities:

$$\begin{aligned}
 I(L; U) &\stackrel{(a)}{\leq} I(L; U | \tilde{Z}) \stackrel{(b)}{\leq} I(\hat{Y}; U | \tilde{Z}) \stackrel{(c)}{\leq} I(\mathcal{A}_n(S_n); U | \tilde{Z}) \\
 &\stackrel{(d)}{\leq} I(\mathcal{A}_n(S_n); S_n)
 \end{aligned} \tag{5.14}$$

Fig. 5.2 presents the conditional independence relationships that hold among the various random variables we have introduced. In the chain of inequalities, (a) follows from the fact that $U \perp\!\!\!\perp \tilde{Z}$; (b), (c) follow from the data-processing inequality; and (d) is shown by Haghifam, Negrea, Khisti, Roy, and Dziugaite [HNKRD20]. Except for the first quantity, $I(L; U)$, each quantity, or some close analogue, has been studied in the literature. For example, $I(\hat{Y}; U | \tilde{Z})$ was studied by Harutyunyan, Raginsky, Ver Steeg, and Galstyan [HRVG21] in the context of CMI with a supersample of size

2n.

We now show that $I(L; U)$ exactly determines the risk (equivalently, expected generalization error) of consistent algorithms for 0–1 loss. For bounded loss functions, $I(L; U)$ upper-bounds the generalization error of arbitrary learning algorithms.

Theorem 5.5.1. *Let \mathcal{D} be a distribution on \mathcal{Z} and let \mathcal{A}_n be a learning algorithm for $n \in \mathbb{N}$ i.i.d. data. For any $[0, 1]$ -bounded loss function,*

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \sqrt{2I(L; U)}.$$

If \mathcal{A}_n is almost surely interpolating, then, for zero-one loss,

$$R_{\mathcal{D}}(\mathcal{A}_n) = \frac{I(L; U)}{\log(n+1)}.$$

Proof. Consider a $[0, 1]$ -bounded loss. For all $i \in [n+1]$, let $\rho_i : [n+1] \rightarrow \mathbb{R}$ be

$$\rho_i(j) = \begin{cases} 1 & \text{if } j = i, \\ -\frac{1}{n} & \text{otherwise.} \end{cases}$$

Let \mathcal{D} , \tilde{Z} , U , L , and \tilde{U} be as in the proof of Theorem 5.2.1. By the Donsker–Varadhan variational formula [BLM13, Prop. 4.15], for all bounded measurable functions h and for all $\lambda \in \mathbb{R}$

$$\begin{aligned} I(L; U) &= \text{KL}(\mathbb{P}(U, L) \parallel \mathbb{P}(L) \otimes \mathbb{P}(\tilde{U})) \\ &\geq \mathbb{P}(U, L)(\lambda h) - \log [(\mathbb{P}(L) \otimes \mathbb{P}(\tilde{U}))(\exp(\lambda h))]. \end{aligned} \tag{5.15}$$

Consider now the function $h : [n+1] \times [0, 1]^{n+1} \rightarrow [-1, 1]$ given by $h(u, l) = \sum_{i=1}^{n+1} \rho_i(u) l_i$. Then

$$\begin{aligned} \mathbb{P}(U, L)(h) &= \mathbb{E}[\mathbb{E}^U[h(U, L)]] \\ &= \mathbb{E}[\mathbb{E}^U[\ell(\mathcal{A}_n(\tilde{Z}_{-U}), \tilde{Z}_U) - \frac{1}{n} \sum_{i \in [n], i \neq U} \ell(\mathcal{A}_n(\tilde{Z}_{-U}), \tilde{Z}_i)]] \\ &= \mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n)) - \hat{R}_{S_n}(\mathcal{A}_n(S_n))] \\ &= \text{EGE}_{\mathcal{D}}(\mathcal{A}_n). \end{aligned} \tag{5.16}$$

Moreover, for all $i \in [n+1]$, we have $\mathbb{E}[\rho_i(\tilde{U})] = 0$. Therefore $\mathbb{E}^L h(\tilde{U}, L) = 0$ since

$\tilde{U} \perp\!\!\!\perp L$. Using Hoeffding's lemma and the fact that $|h| \leq 1$, we obtain

$$\mathbb{P}(L) \otimes \mathbb{P}(\tilde{U})(\exp(\lambda h)) = \mathbb{E}[\mathbb{E}^L \exp(\lambda h(\tilde{U}, L))] \leq \exp(\lambda^2/2). \quad (5.17)$$

By Eqs. (5.15) to (5.17), $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{I(L;U)}{\lambda} + \frac{\lambda}{2}$. Finally, letting $\lambda = \sqrt{2I(L;U)}$, we obtain the stated result.

Now consider 0–1 loss and assume \mathcal{A}_n is interpolating. By the definition of the mutual information, we have $I(L;U) = \text{H}(L) - \text{H}(L|U)$. Let $0_{(i)}$, $i \in \{0, \dots, n+1\}$ be as in the proof of Theorem 5.3.3. Because \mathcal{A}_n is interpolating, the support of L is $\{0_{(i)} | i \in \{0, \dots, n+1\}\}$. For $i > 0$,

$$\mathbb{P}(L = 0_{(i)}) = \frac{1}{n+1} \mathbb{P}(\ell(\mathcal{A}_n(\tilde{Z}_{-i}), \tilde{Z}_i) = 1) = \frac{R_{\mathcal{D}}(\mathcal{A}_n)}{n+1},$$

where we have used the fact that $R_{\mathcal{D}}(\mathcal{A}_n) = \mathbb{P}(\ell(\mathcal{A}_n(\tilde{Z}_{-i}), \tilde{Z}_i) = 1)$ for $i > 0$. Therefore,

$$\begin{aligned} \text{H}(L) &= - \sum_{i=0}^{n+1} \mathbb{P}(L = 0_{(i)}) \log(\mathbb{P}(L = 0_{(i)})) \\ &= -(1 - R_{\mathcal{D}}(\mathcal{A}_n)) \log(1 - R_{\mathcal{D}}(\mathcal{A}_n)) - R_{\mathcal{D}}(\mathcal{A}_n) \log \frac{R_{\mathcal{D}}(\mathcal{A}_n)}{n+1}. \end{aligned} \quad (5.18)$$

Similarly, we have

$$\begin{aligned} \text{H}(L|U) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \text{H}^{U=i}(L) \\ &= -(1 - R_{\mathcal{D}}(\mathcal{A}_n)) \log(1 - R_{\mathcal{D}}(\mathcal{A}_n)) - R_{\mathcal{D}}(\mathcal{A}_n) \log R_{\mathcal{D}}(\mathcal{A}_n). \end{aligned} \quad (5.19)$$

Using Eq. (5.18), Eq. (5.19), and the definition of the mutual information, the stated result follows. \square

Remark 5.5.2 (Proof of Theorem 5.2.5). From Theorem 5.5.1 and the inequality $I(L;U) \leq I(L;U|\tilde{Z})$, we obtain Theorem 5.2.5. \triangleleft

Remark 5.5.3 (Maximal gaps). Using the fact that $I(\mathcal{A}_n(S_n); U|\tilde{Z}) \leq \text{H}(U) \leq \log(n+1)$, Theorem 5.5.1 and Eq. (5.14) imply that, for 0–1 loss and interpolating learning algorithms, we have

$$R_{\mathcal{D}}(\mathcal{A}_n) = \frac{I(L;U)}{\log(n+1)} \leq \frac{I(L;U|\tilde{Z})}{\log(n+1)} \leq \frac{I(\hat{Y}; U|\tilde{Z})}{\log(n+1)} \leq \frac{I(\mathcal{A}_n(S_n); U|\tilde{Z})}{\log(n+1)} \leq 1.$$

Starting from this chain of inequalities, we can investigate the gap between these

different measures of dependency. Of course, the maximal (additive) gap between all these measures is $\log(n+1)$. Is this gap achieved?

- $I(\hat{Y}; U|\tilde{Z})$ versus $I(\mathcal{A}_n(S_n); U|\tilde{Z})$: Note that, for binary classification under 0–1 loss, we have $I(L; U|\tilde{Z}) = I(\hat{Y}; U|\tilde{Z})$. This is due to L being a one-to-one function of \tilde{Z} and \hat{Y} . As shown in Remark 5.3.5, there is a maximal gap between $I(L; U|\tilde{Z})$ and $I(\mathcal{A}_n(S_n); U|\tilde{Z})$. Therefore, there exists a learning scenario in which $I(\mathcal{A}_n(S_n); U|\tilde{Z}) = \Omega(\log(n+1))$ while $I(\hat{Y}; U|\tilde{Z}) \in o(\log(n+1))$.
- $I(L; U|\tilde{Z})$ versus $I(\hat{Y}; U|\tilde{Z})$: Consider a setting where $\mathcal{X} = \mathcal{Y} = [-1, 1]$ and the loss function is $\ell(\hat{y}, (x, y)) = \mathbb{1}[\hat{y}y < 0]$. Let the data distribution be that of $(X, h^*(X))$, where X is uniformly distributed on \mathcal{X} and $h^*(x) = \mathbb{1}[x > 0]$. Consider the learning rule $\mathcal{A}_n(S_n)(x) = y$ where $(x, y) \in S_n$ and $\mathcal{A}_n(S_n)(x) = x$ otherwise. This algorithm is consistent by design and the expected risk of this algorithm is zero. Therefore, Theorem 5.3.3 and Theorem 5.5.1 show that $I(L; U) = I(L; U|\tilde{Z}) = 0$. However, it can be easily seen that $I(\hat{Y}; U|\tilde{Z}) = \log(n+1)$. Thus there exists a learning scenario such that there is a maximal gap between $I(\hat{Y}; U|\tilde{Z})$ and $I(L; U|\tilde{Z})$.
- $I(L; U)$ versus $I(L; U|\tilde{Z})$: Our result in Theorem 5.3.3 shows that the gap between these two quantities cannot be maximal. Also, as mentioned in Remark 5.3.4, for learning algorithms whose expected risk decays polynomially in n to zero, $I(L; U)$ can only be tighter than $I(L; U|\tilde{Z})$ by a constant factor. Note that since $I(L; U) \log(n+1)$ is the risk, any characterization of the gap between $I(L; U)$ and $I(L; U|\tilde{Z})$ is the same as characterization of the gap between $I(L; U|\tilde{Z})$ and the risk. Understanding the exact gap between these two measures is important future work.

◁

Remark 5.5.4. One advantage of working with $I(L; U|\tilde{Z})$ and $I(\hat{Y}; U|\tilde{Z})$, over non-evaluated LOO^eCMI $I(\mathcal{A}_n(S_n); U|\tilde{Z})$, is that the former quantities do not require one to have a parametrization of the set of possible classifiers. Of course, the training data themselves always serve as a “parametrization” for nonrandomized learning rules, but such a parameterization leads to vacuous bounds. (The same roadblocks pertain to plain CMI.) A quintessential example of a setting where natural parametrization may not exist is that of transductive learning algorithms, i.e., ones whose input includes the test input and whose output is the corresponding label prediction. The k -Nearest Neighbor Algorithm is one specific example.

◁

Remark 5.5.5. The leave-one-out CMI framework provides numerous estimates of the expected generalization error based on various measures of information. Which measure should one use to understand generalization? To simplify the discussion, let us focus on the case of interpolating learners under 0–1 loss, in which case expected generalization error is simply risk. Theorem 5.5.1 indicates that $I(L; U)$ is equal to the risk. As such, there appears to be no advantage to studying $I(L; U)$. Of course, once one recognizes that risk is a mutual information, one can invoke information-theoretic results to obtain quantities that may be easier to estimate, such as provided by the bound $I(L; U) \leq I(L; U|\tilde{Z})$.

Even if one can directly bound or compute $I(L; U)$, there may be advantages to studying measures that are never tighter, such as the quantities later in the chain of inequalities in Eq. (5.14). As argued by Dziugaite, Drouin, Neal, Rajkumar, Caballero, L. Wang, Mitliagkas, and Roy [DDNRCWMMR20], if a formally looser quantity controls generalization error (or risk) to a sufficient extent, then this looser quantity provides a more general explanation of the empirical generalization phenomena, as the adequacy of any formally tighter bound is then tautological. That is, when attempting to explain generalization phenomena, use the loosest bound that suffices.

This perspective also suggests that identifying tighter bounds should not be the goal of studying generalization from an information-theoretic perspective. Instead, we seek a rich hierarchy of bounds and an understanding of their interrelationships, so that we can come to understand generalization in specific instances in terms of the level in this hierarchy needed to explain the phenomenon.

Besides the challenge of identifying the right quantity to explain a phenomenon of interest, there are statistical and computational barriers to studying generalization. The measures of information presented here and studied by other authors all depend on the data distribution, which in many interesting settings is not known, other than through a random sample. Even in cases where certain distributions are known, many of these quantities are computationally intractable, without exploiting special structure. There are ways to navigate around these roadblocks. One example is demonstrated by work using “data-dependent estimates” to study generalization in iterative algorithms in deep learning [NHDKR19; HNKRD20; WHGC21]. \triangleleft

Chapter 6

Limitations of Information-Theoretic Generalization Bounds for Gradient Descent Methods in Stochastic Convex Optimization

6.1 Introduction

In this chapter, we uncover limitations of information-theoretic techniques towards analyzing stochastic gradient descent. To do so, we extend existing information-theoretic frameworks for reasoning about generalization to the setting of stochastic convex optimization (SCO) [SSSS09]. Despite the resulting bounds being provably tight, we develop an SCO problem in which the mutual information terms underlying these bounds are too large to demonstrate that subgradient methods [Cau+47; RM51; Bub15] obtain minimax rates. We also consider the introduction of isotropic Gaussian noise to the final iterate and demonstrate a fundamental tradeoff between optimization error and expected generalization error that never yields minimax rates. Our results also cast doubt on the effectiveness of using isotropic Gaussian noise to study subgradient methods in other settings, such as deep learning.

Information-theoretic bounds are, by their nature, distribution- and algorithm-dependent. These bounds have shown some promises: for instance, these key properties enable information-theoretic frameworks to achieve numerically non-vacuous generalization guarantees for stochastic gradient Langevin dynamics (SGLD) with modern deep-learning datasets and architectures [NHDKR19; HNKRD20; LLQ20;

[BCLZ22](#)]. Therefore, it is natural to wonder whether the underlying quantity—mutual information—offers a potentially unifying tool to reason about generalization.

Information-theoretic techniques have long been used to classify the hardness of learning problems in terms of lower bounds on minimax risk. The development of information-theoretic techniques to upper bound minimax risk is a more recent approach. A stream of work has produced a variety of bounds on the generalization error of learning algorithms in terms of the (conditional) mutual information between the training data and certain statistics of the learned predictor. In the case of binary classification, the suprema of certain such bounds match (known) minimax rates, where the suprema runs over data distributions. In the special case of an interpolating classifier achieving zero empirical risk, the risk is shown to equal a certain mutual information term and to be controlled—for polynomial or slower rates—by upper bounds obtained by conditioning [[HMRK22](#)].

Despite these successful applications, much less is known about the optimality or limitations of these techniques beyond the setting of binary classification and 0–1 valued loss. In this work, we turn our attention to stochastic convex optimization (SCO), a well-studied setting with known minimax rates, and look in particular at the analysis of stochastic gradient methods like stochastic gradient descent (SGD). In contrast to learning with a 0–1 valued loss, the minimax excess risk cannot be characterized in terms of uniform convergence of the generalization error [[SSSS10](#)].

To start, we develop a tight information-theoretic bound for SCO problems, analogous to those developed for classification. We focus on the convex–Lipshitz–bounded (CLB) subclass of SCO learning with gradient descent (GD). Our main result demonstrates that despite the bound being tight, it cannot achieve known minimax rates in the CLB setting for GD.

Next, we investigate whether the gap arises due to GD’s deterministic nature: *can we close the gap by introducing randomness?* In other words, can we find a “*surrogate*” algorithm with good information-theoretic generalization guarantees, and such that this surrogate algorithm is close in generalization to the original one? Such an approach was formalized in [[NDR20](#); [SGRS22](#); [BGZV21](#)], and appears frequently in the generalization literature, e.g. [[NDHR21](#); [WM22](#); [HRVG21](#); [LC02](#); [HD21](#); [DR17](#); [DHGAR21](#); [NBS18](#); [ZHBHB20](#); [CNS20](#); [FKMN20](#); [WXW20](#); [PA22](#)]. The most commonly-used surrogate is a “Gaussian surrogate”, which perturbs the output of the algorithm by adding a Gaussian random variable. Surprisingly, we show that the limitations of information-theoretic analyses in the SCO setting are not eliminated even under the Gaussian surrogate.

Our negative results for Gaussian surrogates cast some doubt on their use to study SGD in other settings, such as deep learning. Information-theoretic techniques have shown some promise in this setting. Building off the seminal work of Pensia, Jog, and Loh [P JL18], information-theoretic generalization bounds were shown to yield numerically nonvacuous estimates for stochastic gradient Langevin dynamics (SGLD) when applied to optimizing overparametrized neural networks on nontrivial deep learning classification benchmarks [NHDKR19; HNKRD20; LLQ20; BCLZ22]. And while existing information-theoretic techniques seemingly cannot be applied directly to stochastic gradient methods like SGD itself, Neu, Dziugaite, Haghifam, and Roy [NDHR21] showed how to obtain a (suboptimal) generalization bound for SGD using an information-theoretic bound for a noisy “surrogate” learning algorithm, designed to track the behavior of SGD. Our results explain the suboptimality of this approach and motivate work understanding the power or limitations of other surrogates.

6.1.1 Contributions

1. We prove *tight* generalization bounds based on the input-output mutual information (IOMI) of Russo and J. Zou [RZ16] and Xu and Raginsky [XR17] and the conditional mutual information (CMI) of Steinke and Zakyntinou [SZ20a] for CLB subclasses of SCO problems, as well as their individual sample variations [NHDKR19; HNKRD20; BZV20b; RBTS20; ZTL21] and evaluated CMI [SZ20a]. Our generalization bounds may be of independent interest and can be used to obtain distribution- and algorithm-dependent generalization bounds for SCO problems beyond the worst-case guarantees.
2. We investigate whether we can *directly* analyze the generalization of GD with our information-theoretic generalization bounds. We provide a negative answer to this question by showing that neither the CMI nor IOMI frameworks can properly characterize the excess risk of GD in SCO problems in the minmax setting. We also extend our negative results to the alternative variations of IOMI and CMI, such as evaluated CMI [SZ20a], and individual sample bounds [NHDKR19; HNKRD20; BZV20b; RBTS20; ZTL21].
3. We consider a surrogate algorithm based on a Gaussian perturbation of the final iterate of GD. We show that the generalization of GD can be decomposed as the sum of the generalization of the perturbed final iterate and a residual term that captures the sensitivity of the loss function to perturbations around such

iterate. We consider a favorable setting where the parameters of the surrogate can be tuned based on the data distribution. Nevertheless, we show that there exists a sequence of CLB problems that can be learned with GD but IOMI and CMI frameworks fail to capture learnability in the minimax sense. Our construction is inspired from the ideas by Amir, Koren, and Livni [AKL21] but with a completely different analysis.

4. We complement our results by showing that our construction also implies the failure of high-probability PAC-Bayes bounds in characterizing learnability of the CLB subclass of SCO problems using GD in the minimax sense. In particular, we prove that the *classical* PAC-Bayes bound of McAllester [McA99b] and the recently proposed *conditional* PAC-Bayes bound of Grunwald, Steinke, and Zakyntinou [GSZ21] are *vacuous* in the minimax sense.

6.1.2 Related Work

Recently, there has been a significant interest in understanding whether information-theoretic generalization bounds can characterize worst-case (minimax) rates for certain learning problems. For binary classification, Bassily, Moran, Nachum, Shafer, and Yehudayoff [BMNSY18] and Livni and Moran [LM20] show that the IOMI and classical PAC-Bayes frameworks of [RZ16; XR17; McA99b] provably fail to characterize the learnability of Vapnik–Chervonenkis classes for which we have strong generalization guarantees. Then, Steinke and Zakyntinou [SZ20a], Grunwald, Steinke, and Zakyntinou [GSZ21], and Haghifam, Dziugaite, Moran, and Roy [HDMR21] show that CMI [SZ20a] can be used to establish optimal bounds in the realizable setting. The results of [SZ20a; GSZ21; HDMR21] show that existing IT bounds *characterize* the minimax rates, *without the need for surrogates*. See also [HSG22; PNG22; NB21]. Our work is different from the prior work since we study limitations of information-theoretic generalization bounds in the context of gradient descent methods. Moreover, our results indicate that existing techniques fail to characterize the minimax rates for gradient descent methods in SCO problems. Our findings stand in stark contrast to the success of information-theoretic frameworks in capturing the learnability of VC classes.

6.2 Preliminaries

6.2.1 Probability and Information Theory Notation

Let P, Q be probability measures on a measurable space. For a P -integrable function f , let $P[f] = \int f dP$. When Q is absolutely continuous with respect to P , denoted $Q \ll P$, we write $\frac{dQ}{dP}$ for (an arbitrary version of) the Radon–Nikodym derivative (or density) of Q with respect to P . The *KL divergence* (or *relative entropy*) of Q with respect to P , denoted $\text{KL}(Q \parallel P)$, is defined as $Q[\log \frac{dQ}{dP}]$ when $Q \ll P$ and as infinity otherwise.

For a random element X in some measurable space \mathcal{X} , let $\mathbb{P}[X]$ denote its distribution, which lives in the space $\mathcal{M}_1(\mathcal{X})$ of all probability measures on \mathcal{X} . Given another random element, say Y in \mathcal{Y} , let $\mathbb{P}^Y[X]$ denote the conditional distribution of X given Y (or, more formally, the σ -algebra induced by Y). If X and Y are independent, denoted by $X \perp\!\!\!\perp Y$, we have $\mathbb{P}^Y[X] = \mathbb{P}[X]$ almost surely (a.s.). Moreover, we write $\mathbb{P}^Z[(X, Y)]$ for the conditional distribution of the pair (X, Y) given a random element Z . For an event, say $X \in A$, $\mathbb{P}^Y[X \in A]$ denotes the event’s conditional probability given Y , which is defined to be the conditional expectation of the indicator random variable $\mathbb{1}[X \in A]$ given Y , denoted $\mathbb{E}^Y \mathbb{1}[X \in A]$. By the law of total expectation (a.k.a. chain or tower rule), $\mathbb{E}\mathbb{E}^{\mathcal{F}} = \mathbb{E}$ for any σ -algebra \mathcal{F} .

The *mutual information between X and Y* is $I(X; Y) = \text{KL}(\mathbb{P}[(X, Y)] \parallel \mathbb{P}[X] \otimes \mathbb{P}[Y])$, where \otimes forms the product measure. Then, the *disintegrated mutual information between X and Y given Z* is $I^Z(X; Y) = \text{KL}(\mathbb{P}^Z[(X, Y)] \parallel \mathbb{P}^Z[X] \otimes \mathbb{P}^Z[Y])$, and the conditional mutual information is $I(X; Y|Z) = \mathbb{E}[I^Z(X; Y)]$.

Let $\mu = \mathbb{P}[X]$ and let $\kappa(Y) = \mathbb{P}^Y[X]$ a.s. If X concentrates on a countable set V with counting measure ν , the (*Shannon*) *entropy of X* is $H(X) = -\mu[\log \frac{d\mu}{d\nu}] = -\sum_{x \in V} \mathbb{P}(X = x) \log \mathbb{P}(X = x)$. The *disintegrated entropy of X given Y* is defined by $H^Y(X) = -\kappa(Y)[\log \frac{d\kappa(Y)}{d\nu}]$, while the *conditional entropy of X given Y* is $H(X|Y) = \mathbb{E}[H^Y(X)]$. Note that $H(X|Y) \leq H(X)$ [CT12].

6.2.2 Stochastic Convex Optimization

A *stochastic convex optimization* (SCO) problem is a triple $(\mathcal{W}, \mathcal{Z}, f)$, where $\mathcal{W} \subseteq \mathbb{R}^d$ is a convex set and $f(\cdot, z) : \mathcal{W} \rightarrow \mathbb{R}$ is a convex function for every $z \in \mathcal{Z}$ [SSSS09]. Informally, given an SCO problem $(\mathcal{W}, \mathcal{Z}, f)$, the goal is to find an approximate

minimizer of the *population risk*

$$F_{\mathcal{D}}(w) := \mathbb{E}_{Z \sim \mathcal{D}}[f(w, Z)],$$

given only an i.i.d. sample $S = \{Z_1, \dots, Z_n\}$ drawn from an unknown distribution \mathcal{D} on \mathcal{Z} .

The *empirical risk* of $w \in \mathcal{W}$ on a sample $S \in \mathcal{Z}^n$ is $\hat{F}_{S_n}(w) := \frac{1}{n} \sum_{i \in [n]} f(w, Z_i)$, where $[n]$ denotes the set $\{1, \dots, n\}$. A *learning algorithm* is a sequence $\mathcal{A} = (\mathcal{A}_n)_{n \geq 1}$ such that, for every positive integer n , \mathcal{A}_n maps S_n to a (potentially random) element $W = \mathcal{A}_n(S_n)$ in \mathcal{W} . The *expected generalization error* of \mathcal{A}_n under \mathcal{D} is $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) = \mathbb{E}[F_{\mathcal{D}}(\mathcal{A}(S_n)) - \hat{F}_{S_n}(\mathcal{A}(S_n))]$.

We refer to \mathcal{W} as the *domain*, to its elements as parameters, to elements of \mathcal{Z} as data, and to f as the *loss function*.

Let \mathcal{L} denote the class of all SCO problems. A subclass $\mathcal{C} \subseteq \mathcal{L}$ is *learnable* if, for every desired accuracy $\epsilon > 0$ and all sufficiently large number of samples n ,

$$\underbrace{\sup_{(\mathcal{W}, \mathcal{Z}, f) \in \mathcal{C}} \inf_{\mathcal{A}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \mathbb{E}[F_{\mathcal{D}}(\mathcal{A}_n(S)) - \inf_{w \in \mathcal{W}} F_{\mathcal{D}}(w)]}_{\text{minimax (expected) excess risk}} < \epsilon,$$

where the infimum runs over algorithms.¹

In general, the class \mathcal{L} itself is not learnable [SB14, Chapter 12]. One important family of subclasses of \mathcal{L} which are known to be learnable are the convex–Lipschitz–bounded (CLB) subclasses of SCO problems where, for constants $L, R \in (0, \infty)$, the loss function $f(\cdot, z)$ is L -Lipschitz for all data instances $z \in \mathcal{Z}$, and the domain \mathcal{W} is closed and has finite diameter R [SB14, Chapter 12]. We denote each such class of SCO problems by $\mathcal{C}_{L,R}$. In the remainder of the chapter, we assume, without loss of generality, that each such \mathcal{W} satisfies $\mathcal{W} \subseteq \{w : \|w\|_2 \leq R\}$.

Let W_S^* denote an arbitrary *empirical risk minimizer* (ERM), i.e., an element of $\arg \min_{w \in \mathcal{W}} \hat{F}_{S_n}(w)$. Then, the expected excess risk, $\mathbb{E}[F_{\mathcal{D}}(\mathcal{A}_n(S)) - F_{\mathcal{D}}(w^*)]$, can be written as the sum

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) + \mathbb{E}[\hat{F}_{S_n}(\mathcal{A}_n(S)) - \hat{F}_{S_n}(W_S^*)] + \mathbb{E}[\hat{F}_{S_n}(W_S^*) - F_{\mathcal{D}}(w^*)],$$

of the *expected generalization error*, *optimization error*, and *approximation error*,

¹Note that the initial sup inf is, by skolemization, equivalent to inf sup, where now the algorithm takes as input both a description of the SCO problem $(\mathcal{W}, \mathcal{Z}, f)$ and the data S_n . We have chosen this presentation for simplicity.

respectively.

The third term satisfies $\mathbb{E}[\hat{F}_{S_n}(W_S^*) - F_{\mathcal{D}}(w^*)] = \mathbb{E}[\hat{F}_{S_n}(W_S^*) - \hat{F}_{S_n}(w^*)] \leq 0$ because W_S^* is an ERM for the training set S , and w^* is a constant. Thus, it often suffices to characterize the expected generalization error and optimization error to obtain tight control of the excess risk. For approaches based on iterative optimization, the optimization error can, in many cases, be bounded by a convergence analysis [Bub15]. Therefore, the problem of controlling expected excess risk frequently amounts to controlling the expected generalization error. Nonetheless, there exist scenarios where the excess risk can vanish while the optimization and generalization errors do not, as shown in [KLMS22] for some CLB problems learned with stochastic gradient descent (SGD).

CLB subclasses can be generically learned by suitably tuned instances of (projected) gradient descent (GD), a long studied algorithm [Cau+47; Bub15]: For a convex and compact subset $\mathcal{W} \subseteq \mathbb{R}^d$, let $\Pi_{\mathcal{W}} : \mathbb{R}^d \rightarrow \mathcal{W}$ denote the Euclidean projection operator, given by $\Pi_{\mathcal{W}}(x) = \arg \min_{y \in \mathcal{W}} \|y - x\|_2$. The GD algorithm, $\text{GD} = (\text{GD}_n)_{n \geq 1}$, is initialized at some feasible point $W_0 \in \mathcal{W}$ and then, for some number T of iterations, proceeds to update the parameters iteratively according to $W_{t+1} = \Pi_{\mathcal{W}}(W_t - \eta_t g_t)$, where η_t is a suitably chosen step-size and $g_t \in \partial \hat{F}_{S_n}(W_t)$ is an element of the subdifferential of $\hat{F}_{S_n}(W_t)$. While there are several variants, we will focus on the case where the output of the algorithm is the final iterate, i.e., $\text{GD}_n(S) = W_T$.

6.2.3 Excess Risk of Gradient Descent

For simplicity, we restrict the discussion to GD with a constant step size, i.e., $\eta_t = \eta$ for all iterations $t \in [T]$. We present known generalization and optimization error bounds for the CLB setting.

In [Ora20], the optimization error of the final iterate of GD in the CLB setting is shown to satisfy

$$\sup_{(W, \mathcal{Z}, f) \in \mathcal{C}_{L,R}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \mathbb{E}[\hat{F}_{S_n}(\text{GD}_n(S)) - \hat{F}_{S_n}(W_S^*)] \leq \frac{R^2}{2\eta T} + \frac{(\log(T) + 2)\eta L^2}{2}. \quad (6.1)$$

(See Lemma D.6.6 for a re-statement of this result in the context of this chapter). A similar result also appears in [Zha04, Thm. 5.3]. Recently, Bassily, Feldman, Guzmán,

and Talwar [BFGT20, Thm. 3.2] proved a generalization bound for GD,

$$\sup_{(\mathcal{W}, \mathcal{Z}, f) \in \mathcal{C}_{L,R}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \text{EGE}_{\mathcal{D}}(\text{GD}_n) \leq 4L^2 \sqrt{T} \eta + \frac{4L^2 T \eta}{n}. \quad (6.2)$$

Together, Equations (6.1) and (6.2) yield the following bound on the excess risk,

$$\sup_{(\mathcal{W}, \mathcal{Z}, f) \in \mathcal{C}_{L,R}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \mathbb{E}[\text{F}_{\mathcal{D}}(\text{GD}_n(S)) - \text{F}_{\mathcal{D}}(w^*)] \leq 4L^2 \eta \left(\sqrt{T} + \frac{T}{n} \right) + \frac{R^2}{2\eta T} + \frac{(\log(T) + 2)\eta L^2}{2}. \quad (6.3)$$

For all $\alpha \geq 2$, Equation (6.3) guarantees that GD achieves an excess risk in $\mathcal{O}(LR/\sqrt{n})$ for a number of iterations $T \in \Theta(n^\alpha)$ and a step-size $\eta \in \Theta(R\sqrt{n}/Ln^\alpha)$. This, in fact, is the best achievable excess risk rate for the class $\mathcal{C}_{L,R}$ in the distribution-free setting [Bub15]. In [AKL21; SSK21], it is shown that GD cannot attain this excess risk rate when the number of iterations satisfies $T \in o(n^2)$.

6.3 Main Questions and Overview of the Results

The generalization error guarantee for GD in Eq. (6.2) is obtained using the algorithmic (uniform) stability framework of Bousquet and Elisseeff [BE02]. (Prior work [HRS16] also relied on algorithmic stability.) As shown above, a particular choice of the GD hyperparameters yields expected generalization error in $\mathcal{O}(LR/\sqrt{n})$. In this chapter, we want to understand whether the same rate can be achieved using an information-theoretic framework for generalization. *Are information-theoretic frameworks for generalization expressive enough to accurately estimate the generalization error of GD for SCO?*

We begin by focusing on two frameworks for measuring the information complexity of a learning algorithm: *input-output mutual information* (IOMI [XR17; RZ15; RZ16]) and *conditional mutual information* (CMI [SZ20a]). The IOMI of an algorithm \mathcal{A}_n with respect to a data distribution \mathcal{D} , denoted $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$, is defined to be the mutual information $I(\mathcal{A}_n(S_n); S_n)$ between the training data S_n and the output of the algorithm, $\mathcal{A}_n(S_n)$. In order to define the CMI framework, consider $n \in \mathbb{N}_+$ training data, let $U = (U_1, \dots, U_n) \sim \text{Unif}(\{0, 1\}^n)$, and let $\tilde{S} = (\tilde{Z}_{i,j})_{i \in \{0,1\}, j \in \{1, \dots, n\}} \sim \mathcal{D}^{\otimes (2 \times n)}$ be a $2 \times n$ array of i.i.d. random elements in \mathcal{Z} , independent from U . Then $\tilde{S}_U = (\tilde{Z}_{U_i, i})_{i=1}^n$ has the same distribution as S_n , and so we may assume, w.l.o.g., that $S_n = \tilde{S}_U$ a.s. The CMI of the algorithm \mathcal{A}_n with respect to the data distribution \mathcal{D} , denoted $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$, is defined to be the conditional mutual information $I(\mathcal{A}_n(S_n); U | \tilde{S})$

between $\mathcal{A}_n(S_n)$ and U given \tilde{S} . In the remainder of the section, we write $\text{IM}_{\mathcal{D}}(\mathcal{A}_n)$ to refer to both $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$.

As the first step towards answering our main question, we develop new generalization bounds in both the IOMI and CMI frameworks to handle the CLB subclass of SCO, and show that our upper bounds are tight. Existing information-theoretic generalization bounds often depend on properties of the loss function $f(w, z)$ for fixed $w \in \mathcal{W}$. For instance, the generalization bounds in [XR17; BZV20a; NHDKR19] depend on the tail of the random variable $f(w, Z)$ when $Z \sim \mathcal{D}$. In SCO, we often have no such control, making it impossible to reason about these problems using existing generalization bounds. Instead, in SCO, it is common for loss functions $f(w, z)$ to have regularity for fixed $z \in \mathcal{Z}$. In Theorem 6.4.1, we develop new information-theoretic generalization bounds for the $\mathcal{C}_{L,R}$ subclass, proving that $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathcal{O}(LR\sqrt{\text{IM}_{\mathcal{D}}(\mathcal{A}_n)/n})$. In Theorem 6.4.2, we show that our bound is *tight* up to constants.

Having obtained $\text{IM}_{\mathcal{D}}(\mathcal{A}_n)$ bounds for SCO problems, we ask whether they capture the generalization properties of GD well enough to obtain minimax rates. In Section 6.5, we provide a negative answer to this question, proving that for sufficiently large n

$$\sup_{(\mathcal{W}, \mathcal{Z}, f) \in \mathcal{C}_{L,R}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \text{IM}_{\mathcal{D}}(\text{GD}_n) \in \Omega(n), \quad (6.4)$$

which implies that neither the CMI nor IOMI frameworks can properly characterize minimax excess risk of GD in SCO problems. In Section 6.7, we study variations of IOMI and CMI, such as evaluated CMI [SZ20a] and individual-sample bounds [BZV20b; NHDKR19; HNKRD20; RBTS20; ZTL21]. We find that they also fail to characterize the generalization of GD algorithm.

Since a direct analysis of GD via $\text{IM}_{\mathcal{D}}(\mathcal{A}_n)$ is not possible, we consider a “surrogate” analysis [NDR20], where an excess risk bound is obtained by comparing the risk of GD to a different (surrogate) algorithm, for which one can obtain generalization guarantees. In other words, *is GD “close” to an algorithm with small information complexity?*

We consider commonly used surrogate, whereby one perturbs the final iterate by a Gaussian random variable (see, e.g., [HD21; DR17; WM22; NDHR21; DHGAR21; NBS18]). More formally, let $\mathcal{A}_n(S) = \tilde{W}$, where $\tilde{W} = \Pi_{\mathcal{W}}(W_T + \xi)$, $W_T = \text{GD}_n(S)$, $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$, and $\xi \perp\!\!\!\perp S$. The generalization of GD can be related to the generalization of the Gaussian surrogate \mathcal{A}_n via the inequality

$$\text{EGE}_{\mathcal{D}}(\text{GD}_n) \leq \text{EGE}_{\mathcal{D}}(\mathcal{A}_n) + \mathbb{E}[\Delta_{\sigma}(W_T)] + \mathbb{E}[\hat{\Delta}_{\sigma}(W_T)], \quad \text{where} \quad (6.5)$$

$$\Delta_\sigma(W_T) = \mathbb{E}^S \left[|F_{\mathcal{D}}(\tilde{W}) - F_{\mathcal{D}}(W_T)| \right] \quad \text{and} \quad \hat{\Delta}_\sigma(W_T) = \mathbb{E}^S \left[|\hat{F}_{S_n}(\tilde{W}) - \hat{F}_{S_n}(W_T)| \right] \quad (6.6)$$

are referred to as *residual terms* in the sequel. In Eq. (6.6), the conditional expectations (given S) marginalize over only the (independent) randomness of ξ . Intuitively, the residual terms measure the sensitivity of the population and empirical loss landscapes [NDHR21]. The sensitivity is measured around W_T to perturbations by an isotropic Gaussian random vector with variance σ^2 .

Remark 6.3.1. In Eq. (6.5), one can drop the absolute values from the residual terms, i.e., second and third terms, to obtain

$$\text{EGE}_{\mathcal{D}}(\text{GD}_n) = \text{EGE}_{\mathcal{D}}(\mathcal{A}_n) + \mathbb{E} \left[F_{\mathcal{D}}(W_T) - F_{\mathcal{D}}(\tilde{W}) \right] + \mathbb{E} \left[\hat{F}_{S_n}(\tilde{W}) - \hat{F}_{S_n}(W_T) \right]. \quad (6.7)$$

In this remark, we want to demonstrate how tautologies can arise if one directly studies Eq. (6.7) instead of Eq. (6.5). Consider a surrogate that simply outputs a fixed parameter from \mathcal{W} (independent of the training set). For instance, let $\mathcal{A}_n(S) = 0$ ($\tilde{W} = 0$). Then $\text{IM}_{\mathcal{D}}(\mathcal{A}_n) = 0$ and $\hat{F}_{S_n}(\tilde{W}) = F_{\mathcal{D}}(\tilde{W})$. Therefore, Eq. (6.7) in this case is simplified to

$$\text{EGE}_{\mathcal{D}}(\text{GD}_n) = \mathbb{E} [F_{\mathcal{D}}(W_T)] - \mathbb{E} [\hat{F}_{S_n}(W_T)] = \text{EGE}_{\mathcal{D}}(\text{GD}_n), \quad (6.8)$$

taking us back to the original problem.

Next, we argue that even if we restrict the surrogate algorithm to the case of perturbation by Gaussian random variable, i.e., $\mathcal{A}_n(S) = \tilde{W}$ where $\tilde{W} = \Pi_{\mathcal{W}}(W_T + \xi)$, we get an equally tautological statement from the decomposition in Eq. (6.7). In particular, we claim that letting $\sigma \rightarrow \infty$ takes us back to the original problem. Consider the IOMI framework. Since \mathcal{W} is bounded, we have $\text{Var}(W_T) \in \mathcal{O}(1)$. Then, using [PW19, Thm. 4.6] and the data-processing inequality for mutual information, we obtain $I(\Pi_{\mathcal{W}}(W_T + \xi); S) \leq I(W_T + \xi; S) \leq I(W_T + \xi; W_T) \in \mathcal{O}(1/\sigma^2)$ which tend to 0 as σ diverges. Also, as $\sigma \rightarrow \infty$, $\mathbb{E}[\hat{F}_{S_n}(\tilde{W})] \approx \mathbb{E}[F_{\mathcal{D}}(\tilde{W})]$. Therefore, in the case that $\sigma \rightarrow \infty$, by simplifying Eq. (6.7), we arrive at the same tautology as in Eq. (6.8). Since $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ for any learning problem [HNKRD20, Thm. 2.1], we have the same tautology even if we use the CMI framework. \triangleleft

In this Gaussian surrogate setting, the question of whether $\text{IM}_{\mathcal{D}}(\mathcal{A}_n)$ bounds

characterize $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ is equivalent to asking whether

$$\sup_{(\mathcal{W}, \mathcal{Z}, f) \in \mathcal{C}_{L,R}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \inf_{\sigma \geq 0} \left\{ LR \sqrt{\frac{\text{IM}_{\mathcal{D}}(\mathcal{A}_n)}{n}} + \mathbb{E}[\Delta_{\sigma}(W_T)] + \mathbb{E}[\hat{\Delta}_{\sigma}(W_T)] \right\} \stackrel{?}{\in} \Theta\left(\frac{LR}{\sqrt{n}}\right). \quad (6.9)$$

Answering this amounts to answering whether one can choose a value of σ *with a full knowledge of the SCO problem and the data distribution*, such that the perturbed GD algorithm achieves the optimal rate via the generalization bound appearing in Theorem 6.4.1. Alternatively, one can ask whether one can show, using the perturbation idea, that GD learns the subclass $\mathcal{C}_{L,R}$ *even with an arbitrary slow rate*, i.e., whether or not the LHS of Eq. (6.9) converges to zero as the number of the training samples diverges.

In order to gain insight on the Gaussian surrogate, consider extreme values of the variance of the perturbations. Setting $\sigma = 0$ corresponds to a direct analysis of GD, and the result in Eq. (6.4) shows we cannot prove learnability using existing frameworks. At the other extreme, one can show that $\text{IM}_{\mathcal{D}}(\mathcal{A}_n) \rightarrow 0$ as $\sigma \rightarrow \infty$, leaving us with a bound in terms of the sum of the residual terms alone. As the distance between \tilde{W} and W_T is maximal under such a perturbation, the sum of the residual terms is in $\Omega(1)$, once again failing to establish learnability. The idea behind introducing the surrogate algorithm \mathcal{A} and adjusting the value of σ is that it allows one to conceptually interpolate between these two extreme points in order to find an optimal bound on GD's generalization error.

Nevertheless, for the perturbed GD, we prove a negative result showing that

$$\sup_{(\mathcal{W}, \mathcal{Z}, f) \in \mathcal{C}_{L,R}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \inf_{\sigma \geq 0} \left\{ LR \sqrt{\frac{\text{IM}_{\mathcal{D}}(\mathcal{A}_n)}{n}} + \mathbb{E}[\Delta_{\sigma}(W_T)] + \mathbb{E}[\hat{\Delta}_{\sigma}(W_T)] \right\} \in \Omega(1). \quad (6.10)$$

Note that our negative result holds even if the perturbation's variance is allowed to depend on the data distribution \mathcal{D} and the SCO problem $(\mathcal{W}, \mathcal{Z}, f)$. While the distribution is unknown, the surrogate algorithm is a theoretical device and can be chosen with full knowledge of the data distribution to achieve the tightest possible bound. As such, we must control also the infimum.

In Section 6.6, we extend our results to PAC-Bayes bounds, which provide tail bounds on the generalization error of GD, with respect to the randomness in the data. A similar surrogate decomposition as in Eq. (6.5) relates *disintegrated* generalization

of GD to the generalization of \mathcal{A}_n via

$$\mathbb{E}^S \left[F_{\mathcal{D}}(W_T) - \hat{F}_{S_n}(W_T) \right] \leq \mathbb{E}^S \left[F_{\mathcal{D}}(\tilde{W}) - \hat{F}_{S_n}(\tilde{W}) \right] + \hat{\Delta}_{\sigma}(W_T) + \Delta_{\sigma}(W_T), \quad (6.11)$$

where $\hat{\Delta}_{\sigma}(W_T)$ and $\Delta_{\sigma}(W_T)$ are defined in Eq. (6.6). The first term on the RHS of Eq. (6.11) can be analyzed using PAC-Bayes frameworks (see, e.g., [LC02; HD21; DR17; DHGAR21; NBS18; CNS20; FKMN20; WXW20]). In Section 6.6, we show that, in the minimax sense, the classical and conditional PAC-Bayes frameworks of McAllester [McA99b] and Grunwald, Steinke, and Zakyntinou [GSZ21] provide a vacuous characterization of the RHS of Eq. (6.11) for all values of σ .

6.4 Information-Theoretic Generalization Bounds for the CLB setting

In SCO problems, generalization bounds for gradient methods can be obtained using the uniform stability framework [HRS16; BFGT20; FV18; FV19]. This framework provides an algorithm-dependent approach that has been used to obtain relatively strong generalization bounds for several convex optimization algorithms in the distribution-free setting. In this section, we extend the CMI and IOMI frameworks to the CLB setting and provide *algorithm-* and *distribution-* dependent generalization bounds.

Theorem 6.4.1. *Let $n \in \mathbb{N}$, $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be a data distribution, and $S \sim \mathcal{D}^{\otimes n}$. Consider an SCO problem $(f, \mathcal{W}, \mathcal{Z}) \in \mathcal{C}_{L,R}$. Then, for every learning algorithm \mathcal{A}_n such that $\mathcal{A}_n(S) \in \mathcal{W}$ a.s.,*

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq LR \sqrt{\frac{2\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}} \quad \text{and} \quad \text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq LR \sqrt{\frac{8\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}}.$$

The proof, based on [RBTS21], is in Appendix D.1. To better contextualize our generalization bounds in Theorem 6.4.1, we study their tightness. For the trivial case where the output of a learning algorithm is independent of the training set, the bounds in Theorem 6.4.1 are tight. The theorem below states that the bounds are tight even when the learning algorithm depends on the training set.

Theorem 6.4.2. *For every $n \in \mathbb{N}$, $L \in \mathbb{R}_+$, $R \in \mathbb{R}_+$, there exists an SCO problem $(f, \mathcal{W}, \mathcal{Z}) \in \mathcal{C}_{L,R}$, a data distribution \mathcal{D} over \mathcal{Z} , and a learning algorithm $\mathcal{A} = (\mathcal{A}_n)_{n \geq 1} \in \mathcal{W}$ such that: (i) the expected generalization error of \mathcal{A}_n satisfies $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \geq LR/\sqrt{2n}$, and (ii) the upper bounds from Theorem 6.4.1 are*

$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq LR\sqrt{2}/\sqrt{n}$ and $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq LR\sqrt{8}/\sqrt{n}$, respectively.

See Appendix D.2 for the proof, which is inspired by [Ora19, Sec. 5.1]. Theorem 6.4.2 shows that there exists a learning algorithm in the CLB setting for which the bounds Theorem 6.4.1 is tight. This implies that the bound in Theorem 6.4.1 cannot be *uniformly* improved for every learning algorithm in the CLB setting. Note, however, that there may exist a tighter bound for some learning algorithms.

6.5 Failure of Information-Theoretic Bounds for GD in the CLB Setting

An important feature of GD for SCO problems is that the sample complexity is *dimension-independent*: For every SCO problem in $\mathcal{C}_{L,R}$, if L and R do not grow with the parameters' (ambient) dimension, one needs $\mathcal{O}(1/\epsilon^2)$ samples to reach ϵ expected excess risk using GD, regardless of the dimension. In this section, we exploit this property to show that the *distribution-free learnability* of SCO in the CLB setting using GD cannot be explained using the IOMI or CMI frameworks.

Let $\text{GD}(S, \eta, T)$ denote the output of gradient descent, training on the training set S with learning rate η for T iterations, starting from a zero initialization.

Theorem 6.5.1. *Let $n \in \mathbb{N}$, $T_{(n)} = 2n^2$, $\eta_{(n)} = \frac{1}{n\sqrt{5n}}$, and $d_{(n)} = 3T_{(n)}/4$. Then, there exists a universal constant $N^* \in \mathbb{N}$ such that for every $n \geq N^*$, there exist a sequence of SCO problems $\{(f_{(n)}, \mathcal{W}_{(n)}, \mathcal{Z}_{(n)}) \in \mathcal{C}_{4,1}\}_{n \in \mathbb{N}}$ where $\mathcal{W}_{(n)}, \mathcal{Z}_{(n)} \in \mathbb{R}^{d_{(n)}}$, and a data distribution $\mathcal{D}_{(n)}$ over $\mathcal{Z}_{(n)}$ such that the following holds: For $S \sim \mathcal{D}_{(n)}^{\otimes n}$, let $W_T = \text{GD}(S, \eta_{(n)}, T_{(n)})$ and $\mathcal{A}_n(S) = \Pi_{\mathcal{W}}(W_T + \xi)$, where $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_{(n)}})$. Then, there exists $\text{var}_n^* > 0$ such that if $\sigma^2 \leq \text{var}_n^*$, then $\text{IOMI}_{\mathcal{D}_{(n)}}(\mathcal{A}_n) \in \Omega(n^3)$ and $\text{CMI}_{\mathcal{D}_{(n)}}(\mathcal{A}_n) \in \Omega(n)$. Also, if $\sigma^2 > \text{var}_n^*$, then $\mathbb{E}[\hat{\Delta}_\sigma(W_T)] + \mathbb{E}[\Delta_\sigma(W_T)] \in \Omega(1)$. As a result,*

$$\inf_{\sigma \geq 0} \left\{ \sqrt{\frac{\min\{2\text{IOMI}_{\mathcal{D}_{(n)}}(\mathcal{A}_n), 8\text{CMI}_{\mathcal{D}_{(n)}}(\mathcal{A}_n)\}}{n}} + \mathbb{E}[\hat{\Delta}_\sigma(W_T)] + \mathbb{E}[\Delta_\sigma(W_T)] \right\} \in \Omega(1),$$

while the generalization error of GD satisfies $\mathbb{E}[|F_{\mathcal{D}_{(n)}}(W_T) - \hat{F}_{S_n}(W_T)|] \in \mathcal{O}(1/\sqrt{n})$.

Proof. Here, we provide an overview of the proof. The formal proof can be found in Appendix D.3. Our construction is inspired from the construction in Amir, Koren, and Livni [AKL21].

• **Construction and Dynamics of GD:** Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{0, 1\}^d$. Let the data distribution on input be $(\text{Ber}(1/2))^{\otimes d}$, i.e., each coordinate is drawn independently and uniformly at random from $\text{Ber}(1/2)$. Thus, the training set $S \in \{0, 1\}^{n \times d}$ is a matrix whose elements are drawn i.i.d. from $\text{Ber}(1/2)$. Let λ be a sufficiently small constant, and \mathcal{W} be a ball of radius one in \mathbb{R}^d . We consider the following loss function $f : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}$, $f(w, z) = \sum_{i=1}^d z(i)w(i)^2 + \lambda \langle w, z \rangle + \max\{\max_{i \in [d]} \{w(i)\}, 0\}$. We show that this function is convex and 4-Lipschitz. As a result, the problem is in the CLB subclass. Next, we demonstrate that when the dimension is $d = 3T2^n/4$, there are many columns in S such that *all* the entries are zero. Following Amir, Koren, and Livni [AKL21], we refer to such columns as *bad coordinates*. Let $\mathbf{B} \in \{0, 1\}^d$ be a vector whose i -th coordinate is one if and only if i is a bad coordinate. We show that, with high probability, the number of bad coordinates is between $T/2$ and T . The result emerges from the observation that the dynamics of GD along the bad coordinates are completely *different* compared to the good coordinates, therefore “revealing” which coordinates are bad. To see this, consider the empirical risk $\hat{F}_{S_n}(w) = \sum_{i=1}^d \hat{\mu}(i)w(i)^2 + \lambda \langle \hat{\mu}, w \rangle + \max\{\max_{i \in [d]} \{w(i)\}, 0\}$, where for $i \in [d]$, $\hat{\mu}(i) = \frac{1}{n} \sum_{j=1}^n z_j(i) \in [0, 1]$ is the empirical mean of the points in the i -th column of S . By the definition of the bad coordinates we can write $\hat{F}_{S_n}(w) = \sum_{i \in \{i: \mathbf{B}[i]=0\}} \hat{\mu}(i)w(i)^2 + \lambda \sum_{i \in \{i: \mathbf{B}[i]=0\}} \hat{\mu}(i)w(i) + \max\{\max_{i \in [d]} \{w(i)\}, 0\}$. Note that the third term is not differentiable. We consider a specific first-order oracle proposed in [AKL21; BFGT20]. We show that to analyze the dynamics of GD for good coordinates, we only need to consider the first two terms. For good coordinates, the main observation here is that because of the *norm-like* penalty from the first term, $|W_T(i)|$ is small. In contrast, for the bad coordinates the gradient that comes from the third term pushes $W_T(i)$ away from zero; in particular, for the bad coordinates we have $|W_T(i)| = \eta$ under the event $T/2 \leq \|\mathbf{B}\|_0 \leq T$. The other key property used in the proof is that $\|W_T\| \in \Theta(1/\sqrt{n})$ with high probability, meaning that the final iterate of GD is close to the origin.

• **Lower Bound on the Residual Term:** First, we prove that if $\sigma^2 \in \Omega(1/d)$, then the residual term is large. Recall that $\|W_T\| \in \Theta(1/\sqrt{n})$, and $\tilde{W} = \Pi_{\mathcal{W}}(W_T + \xi)$. Consider $\mathbb{E} \left[|F_{\mathcal{D}}(\tilde{W}) - F_{\mathcal{D}}(W_T)| \right]$, where the population risk is given by $F_{\mathcal{D}}(w) = 1/2 \|w\|^2 + \lambda/2 \sum_{i=1}^d w(i) + \max\{\max_{i \in [d]} \{w(i)\}, 0\}$. Using concentration inequalities for Gaussian random variables, we show that $\|\tilde{W}\|^2 \approx \min\{\sigma^2 d + o(1), 1\}$, while $\|W_T\|^2 \in o(1)$. Using this argument we show that unless $\sigma^2 \in \mathcal{O}(1/d)$, the residual term grows with n . Since d is exponentially large in n , we conclude that the variance

of noise has to satisfy $\sigma^2 \in \mathcal{O}(2^{-n})$.

- Lower Bound on $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$:** Here we show that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \in \Omega(n)$, which implies that $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n) \in \Omega(n)$, since $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ [HNKRD20, Thm. 2.1]. In Appendix D.3, we prove the stronger result $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n) \in \Omega(n^3)$. Step one is to establish that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq n - (\text{H}(\mathbf{B}|\tilde{W}, \tilde{S}) + \text{H}(U|\tilde{W}, \tilde{S}, \mathbf{B})) \geq n - (\text{H}(\mathbf{B}|\tilde{W}) + \text{H}(U|\tilde{S}, \mathbf{B}))$ using standard properties of mutual information. Next, we seek to upper bound $\text{H}(\mathbf{B}|\tilde{W})$ and $\text{H}(U|\tilde{S}, \mathbf{B})$. We do so using *Fano's inequality* (Lemma D.6.4) but in a way that differs from its conventional use. The intuition behind using Fano's inequality is as follows: if there exists an estimator that can be used to predict \mathbf{B} using \tilde{W} , then the conditional entropy $\text{H}(\mathbf{B}|\tilde{W}_T)$ is small. The same also holds for predicting U using $\tilde{W}, \tilde{S}, \mathbf{B}$. The core of the proof then rests on designing two estimators: one for estimating \mathbf{B} using \tilde{W} , and another one for estimating U using \tilde{S} and \mathbf{B} . We construct explicit estimators for each, and demonstrate that their probability of error is small. Thus Fano's inequality implies that the entropy terms of interest are small. To construct the first estimator, we use two important properties: (i) the variance of noise satisfies $\sigma^2 \in \mathcal{O}(2^{-n})$, and (ii) for the good coordinates $|W_T(i)|$ is very small, while for the bad coordinates we have $|W_T(i)| \in \Theta(n^{-1.5})$. The proposed estimator is based on comparing $|\tilde{W}(i)|$ with a threshold. We show that σ^2 is much smaller than $|W_T(i)|$ for the bad coordinates. As a result, the Gaussian noise does not *perturb* the bad coordinates significantly. Thus, the error probability of this estimator can be arbitrarily small as n diverges. For constructing the second estimator, remember that: (i) by definition, in each column of \tilde{S} exactly one sample is chosen for the training set, and (ii) by the definition of the bad coordinates, we know that if $i \in [d]$ is a bad coordinate, then for all $Z \in S$, we have $Z(i) = 0$. Therefore, in every column of the supersample, either one or both of the samples have *zeros in all of the bad coordinates*. Our proposed estimator is as follows: whenever there is only one sample, the estimator can perfectly recover U for that column. In the case of two samples, the estimator makes a random guess. We show that the probability that there are two samples in a column such that both have zeros in all of the bad coordinates is $\Theta(2^{-n^2})$. Therefore, the estimator makes an error with small probability.

□

Remark 6.5.2. The sequence of SCO problems that witnesses that lower bound for the IOMI and CMI frameworks is *the same*. Hence, a tight generalization bound

cannot be achieved for every SCO problem by considering the best framework for that problem out of the IOMI and CMI frameworks. \triangleleft

Remark 6.5.3. Equation (6.3) provides a general result for the excess risk guarantee of GD for every number of iterations T and the step size η . GD obtains the excess risk and the generalization error guarantees of $\mathcal{O}(\frac{LR}{\sqrt{n}})$ by setting $T \in \Theta(n^\alpha)$ and $\eta \in \Theta(\frac{R\sqrt{n}}{Ln^\alpha})$ for every $\alpha \geq 2$. In Theorem 6.5.1, we state the results only for $\alpha = 2$. However, the same construction can be used to prove the lower bounds in Theorem 6.5.1 for every $\alpha \geq 2$. This observation shows a stronger failure: for every parameter setting under which GD attains the optimal excess risk, the upper bound in Eq. (6.9) does not even converge to zero, i.e., $\Omega(1)$. \triangleleft

Remark 6.5.4. A notable property of the construction in Theorem 6.5.1 is that the Lipschitz constant of the loss function and the diameter of \mathcal{W} do not grow with dimension. By a simple scaling, our result in Theorem 6.5.1 implies the lower bound stated in Eq. (6.10). \triangleleft

6.6 Implications for PAC-Bayes Bounds

In this section, we show that our construction that witnesses the lower bounds in Theorem 6.5.1 reveals a limitation of PAC-Bayes bounds for learning SCO problems with GD. Using PAC-Bayes bounds to analyze the generalization of gradient methods via the surrogate algorithm that perturbs the final weight with a Gaussian random variable is a prevailing method in the literature [LC02; HD21; DR17; DR18b; DHGAR21; NBS18; CNS20; FKMN20; WXW20]. This approach leads to non-vacuous estimates of the generalization gap for non-convex problems such as training modern deep learning models. However, we show that it fails for the CLB subclass of SCO problems.

We consider a classical PAC-Bayes bound [McA99a] and a recently-proposed conditional PAC-Bayes bound [GSZ21]. The main difference between the two is the *measure of complexity* that characterizes generalization. We can represent an algorithm \mathcal{A}_n with a posterior distribution $Q : \mathcal{Z}^n \rightarrow \mathcal{M}_1(\mathcal{W})$. A complexity measure appearing in classical PAC-Bayes bounds is $C_{\text{clas}}(n) = \text{KL}(Q(S) \parallel \mathbb{E}[Q(S)])$. The conditional PAC-Bayes bound relies on some additional structure. Let $S = (Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$ and $S' = (Z'_1, \dots, Z'_n) \sim \mathcal{D}^{\otimes n}$ such that $S \perp\!\!\!\perp S'$. For every $u = (u_1, \dots, u_n) \in \{0, 1\}^n$, define $\tilde{S}_u = ((1 - u_1)Z_1 + u_1Z'_1, \dots, (1 - u_n)Z_n + u_nZ'_n)$. The complexity measure for the conditional-PAC Bayes bound [GSZ21] is $C_{\text{cond}}(n) = \mathbb{E}^S[\text{KL}(Q(S) \parallel 2^{-n} \sum_{u \in \{0, 1\}^n} Q(\tilde{S}_u))]$. Next we present the known results that relate these complexity measures to the generalization gap.

Theorem 6.6.1 ([McA99a; GSZ21]). *Let $S \sim \mathcal{D}^{\otimes n}$, $\delta \in (0, 1)$, $L, R \in \mathbb{R}_+$. Assume that the range of the loss function f lies in $[-LR, LR]$. Then, with probability at least $(1 - \delta)$ (over the choice of $S \sim \mathcal{D}^{\otimes n}$) for any posterior distribution $Q : \mathcal{Z}^n \rightarrow \mathcal{M}_1(\mathcal{W})$ with $W \sim Q(S)$,*

$$\mathbb{E}^S \left[F_{\mathcal{D}}(W) - \hat{F}_{S_n}(W) \right] \in \mathcal{O} \left(LR \left(\frac{\min\{C_{\text{cond}}(n), C_{\text{clas}}(n)\} + \log(n/\delta)}{n} \right)^{\frac{1}{2}} \right).$$

Let $\text{complexity}(n)$ denote either $C_{\text{clas}}(n)$ or $C_{\text{cond}}(n)$. Note that $\text{complexity}(n)$ is a S -measurable random variable. We next present our main result of this section showing the failure of PAC-Bayes bounds for learning SCO with GD.

Theorem 6.6.2. *Let $n \in \mathbb{N}$, $T_{(n)} = 2n^2$, $\eta_{(n)} = \frac{1}{n\sqrt{5n}}$, $d_{(n)} = 3T_{(n)}/4$, and $N^* \in \mathbb{N}$ be a universal constant. Then, there exists $\omega \in (0, 1)$, a sequence of SCO problems $\{(f_{(n)}, \mathcal{W}_{(n)}, \mathcal{Z}_{(n)}) \in \mathcal{C}_{4,1}\}_{n \in \mathbb{N}}$ where $\mathcal{W}_{(n)}, \mathcal{Z}_{(n)} \in \mathbb{R}^{d_{(n)}}$, and a data distribution $\mathcal{D}_{(n)}$ over $\mathcal{Z}_{(n)}$ such that the following holds for all $n \geq N^*$: For $S \sim \mathcal{D}_{(n)}^{\otimes n}$, let $W_T = \text{GD}(S, \eta_{(n)}, T_{(n)})$ and $\mathcal{A}_n(S) = \Pi_{\mathcal{W}}(W_T + \xi)$, where $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_{(n)}})$. Then, for every $0 < \delta < 1 - \omega$, with probability at least $1 - \delta - \omega$ over $S \sim \mathcal{D}_{(n)}^{\otimes n}$,*

$$\inf_{\sigma \geq 0} \max \left\{ \sqrt{\frac{\text{complexity}(n) + \log(n/\delta)}{n}}, \hat{\Delta}_{\sigma}(W_T) + \Delta_{\sigma}(W_T) \right\} \in \Omega(1).$$

This result implies that a PAC-Bayes bound for the surrogate from Eq. (6.11) yields a *vacuous* generalization bound with *constant probability*, i.e., independent of n .

Remark 6.6.3. The complexity term in PAC-Bayes bounds generally takes the form $\text{KL}(Q(S) \parallel P)$ for some element $P \in \mathcal{M}_1(\mathcal{W})$. The choice here, $P = \mathbb{E}[Q(S)]$, minimizes the complexity term in expectation. Whether other choices might yield tighter high probability bounds is left open. \triangleleft

6.7 Failure of Information-Theoretic Alternatives to the IOMI and CMI Frameworks

In the previous sections, we showed how the IOMI and the CMI frameworks and their high-probability counterparts fail to characterize the behavior of GD in the CLB setting, even when they are strengthened with a surrogate analysis. In this section, we consider other alternatives and reinforcements of these two frameworks and show that

they also fail to characterize the behavior of SCO problems in the CLB setting, albeit without considering any potential strengthening with a surrogate analysis. First, we introduce these alternatives and their motivation, then we adapt them to the CLB setting, and finally we show their failure.

6.7.1 Information-Theoretic Alternatives to the IOMI and CMI Frameworks

The IOMI and CMI frameworks are attractive due to algorithm- and distribution-dependence. Nevertheless, they come with some drawbacks.

- 1 The IOMI may be infinite and the CMI may be $\Omega(n)$ for a variety of learning scenarios, e.g., deterministic algorithms.
- 2 IOMI and CMI may capture unnecessary information. Note that we can write $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n) = \sum_{i=1}^n I(\mathcal{A}(S); Z_i) + I(Z^{i-1}, Z_i | \mathcal{A}(S))$, where $Z^{i-1} := (Z_1, \dots, Z_{i-1})$. It is clear from this decomposition that IOMI not only captures the information that the output contains about individual samples, but also captures the “artificial” dependencies among the samples, given the algorithm’s output. The latter is not predictive of the generalization performance of the algorithm [BZV20b]. An analogous problem arises in CMI, which includes the dependence of the indices and the samples after observing the algorithm’s output [RBTS20].

These problems can be avoided with an *individual-sample* bound proposed in [BZV20b], replacing $I(\mathcal{A}(S); S)$ with the average of $I(\mathcal{A}(S); Z_i)$ for all $i \in [n]$. This bound takes into account the information the output of the algorithm captures about each *individual* sample Z_i , disregarding the generated dependency between the samples after observing the said output. Similarly, the individual-sample bound from [RBTS20; ZTL21] considers the information the output of the algorithm captures about each individual index U_i , disregarding the dependency between the indices and the samples. These bounds are adapted to the CLB setting in the following theorem. The proof is in Appendix D.1.

Theorem 6.7.1. *Let $n \in \mathbb{N}$, $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be a data distribution, and $S \sim \mathcal{D}^{\otimes n}$. Consider an SCO problem $(f, \mathcal{W}, \mathcal{Z}) \in \mathcal{C}_{L,R}$. Then, for every learning algorithm \mathcal{A}_n such that $\mathcal{A}_n(S) \in \mathcal{W}$ a.s., we have $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{LR}{n} \sum_{i=1}^n \sqrt{2I(\mathcal{A}(S); Z_i)}$, and $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{2LR}{n} \sum_{i=1}^n \sqrt{2I(\mathcal{A}(S); U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i})}$.*

Remark 6.7.2. As mentioned above, the individual-sample alternatives to the IOMI and CMI are tighter than the IOMI and CMI. This may be seen by [BZV20b, Prop. 2] and [RBTS20, Lemma 3] or [ZTL21, Lemma 2], where we have that

$$\frac{1}{n} \sum_{i=1}^n \sqrt{2I(W; Z_i)} \leq \sqrt{\frac{2\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \sqrt{2I(W; U_i | \tilde{Z}_{1,i}, \tilde{Z}_{2,i})} \leq \sqrt{\frac{2\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}}.$$

Therefore, Theorem 6.7.1 implies Theorem 6.4.1. Moreover, we have that $I(\mathcal{A}(S); U_i | \tilde{Z}_{1,i}, \tilde{Z}_{2,i}) \leq I(\mathcal{A}(S), Z_i)$ [RBTS21, App. D.2.3]. \triangleleft

Another drawback of IOMI and CMI frameworks is the following:

- 3 Both the IOMI and CMI of an algorithm depend on the joint distribution of the algorithm's output and other variables. In contrast, generalization error depends on the algorithm's output only through the losses it incurs. Therefore, it is possible to increase both the IOMI and the CMI by *embedding* information about the training set in the output of a learning algorithm without affecting the algorithm's statistical properties [LM20; BMNSY18].

Steinke and Zakynthinou [SZ20a] propose an alternative framework, *evaluated CMI*, that considers the information about the data captured by the *incurred loss* rather than the output itself.

Definition 6.7.3 (Evaluated CMI, [SZ20a, Sec. 6.2.2]). Let $n \in \mathbb{N}$. Let the supersample \tilde{S} and indices U be as defined in Section 6.3. Let $S = (Z_{U_i,i})_{i \in [n]}$, and $F \in \mathbb{R}^{2 \times n}$ be the array with entries $F_{v,i} = f(\mathcal{A}_n(S), Z_{v,i})$ for $v \in \{0, 1\}$, $i \in [n]$. The *evaluated conditional mutual information of \mathcal{A} with respect to \mathcal{D}* , denoted by $e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n))$, is the conditional mutual information $I(F; U | \tilde{S})$.

Haghifam, Dziugaite, Moran, and Roy [HDMR21] show that the eCMI can provide a sharp characterization of generalization for the realizable setting and 0–1 losses. Below, we state a bound for the CLB setting based on eCMI. The proof can be found in Appendix D.1.

Theorem 6.7.4. *Consider an SCO problem $(f, \mathcal{W}, \mathcal{Z}) \in \mathcal{C}_{L,R}$. Then, for every learning algorithm \mathcal{A}_n such that $\mathcal{A}_n(S) \in \mathcal{W}$ a.s., we have $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq LR \sqrt{\frac{8e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n))}{n}}$.*

Remark 6.7.5. Similarly to Remark 6.7.2, note that the evaluated version of the CMI is tighter than the CMI itself, i.e., $e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n)) \leq \text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ [SZ20a]. \triangleleft

Remark 6.7.6. Since these alternatives to the IOMI and CMI are tighter than the IOMI and CMI themselves (cf. Remark 6.7.2 and Remark 6.7.5), the adaptation of these bounds to the CLB setting (Theorems 6.7.1 and 6.7.4) are also tight in the sense of Theorem 6.4.2. \triangleleft

Data-dependent alternatives and functional CMI

Negrea, Haghifam, Dziugaite, Khisti, and Roy [NHDKR19] and Haghifam, Negrea, Khisti, Roy, and Dziugaite [HNKRD20] introduced data-dependent alternatives to the IOMI and CMI frameworks that resulted in numerically non-vacuous generalization guarantees for stochastic gradient Langevin dynamics (SGLD) and its full-batch counterpart for modern deep-learning datasets and architecture. These bounds can also be adapted to the CLB setting by replicating Theorem 6.7.1 (i) considering [RBTS21, Thm. 2] instead of [RBTS21, Thm. 1] and noting that $\mathbb{E}[\text{KL}(\mathbb{P}^S[W] \parallel \mathbb{P}^{S \setminus Z_i}[W])] = I(W; Z_i | S \setminus Z_i)$; and (ii) considering [RBTS21, Thm. 4] instead of [RBTS21, Thm. 3] and noting that $\mathbb{E}[\text{KL}(\mathbb{P}^{\tilde{S}, U}[W] \parallel \mathbb{P}^{\tilde{S}, U \setminus U_i}[W])] = I(W; U_i | \tilde{S}, U \setminus U_i)$. This would yield the data-dependent EGE bounds

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{LR}{n} \sum_{i=1}^n \sqrt{2I(\mathcal{A}(S); Z_i | S \setminus Z_i)} \quad \text{and} \quad (6.12)$$

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{2LR}{n} \sum_{i=1}^n \sqrt{I(\mathcal{A}(S), U_i | \tilde{S}, U \setminus U_i)}. \quad (6.13)$$

Both Eq. (6.12) and Eq. (6.13) are looser than the bounds in Theorem 6.7.1, by similar arguments to those in Remark 6.7.2 [RBTS20, App. J].

Harutyunyan, Raginsky, Ver Steeg, and Galstyan [HRVG21] introduced an alternative to the CMI for supervised learning problems that yield bounds that can be experimentally computed and are non-vacuous. However, by the data processing inequality we have that this notion is looser than the evaluated CMI.

6.7.2 Failure of the Alternatives

We demonstrate that the individual sample and evaluated versions of CMI still fail in the CLB setting. Based on the relative tightness of these alternative frameworks (see Remark 6.7.2 and Remark 6.7.5), showing their failure implies failure of all the aforementioned alternatives to the IOMI and CMI frameworks. In fact, it also proves the failure of (i) the data-dependent bounds from [NHDKR19] and [HNKRD20], and (ii) functional-CMI of [HRVG21], adapted to the CLB setting (c.f. Section 6.7.1).

The following theorem states that the *distribution-free learnability* of GD cannot be *directly* proved using any of the alternatives to the IOMI and CMI framework described above.

Theorem 6.7.7. *Let $n \in \mathbb{N}$, $T_{(n)} = n^2$, $\eta_{(n)} = \frac{1}{n\sqrt{n}}$, and $d_{(n)} = 2n^2$. Then, for every $n \geq 1$, there exists a sequence of SCO problems $\{(f_{(n)}, \mathcal{W}_{(n)}, \mathcal{Z}_{(n)}) \in \mathcal{C}_{1,1}\}_{n \in \mathbb{N}}$ where $\mathcal{W}_{(n)}, \mathcal{Z}_{(n)} \in \mathbb{R}^{d_{(n)}}$, and data distribution $\mathcal{D}_{(n)}$ over $\mathcal{Z}_{(n)}$ such that the following holds: Let $W_T = GD(S, \eta_{(n)}, T_{(n)})$. Then, $e\text{CMI}_{\mathcal{D}_{(n)}}(f(\mathcal{A}_n)) \in \Omega(n)$, and $\sum_{i=1}^n \sqrt{2I(\mathcal{A}(S); U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i})} \in \Omega(n)$, while the generalization error of GD satisfies $\mathbb{E}[|F_{\mathcal{D}_{(n)}}(W_T) - \hat{F}_{S_n}(W_T)|] \in \mathcal{O}(1/\sqrt{n})$.*

Proof. Here, we provide an overview of the proof. The formal proof can be found in Appendix D.5. Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{\mathbf{e}(i) : i \in [d]\}$, where $\mathbf{e}(i) = (0, \dots, 0, 1, 0, \dots, 0)$ with a 1 at the i -th coordinate and $\|\mathbf{e}(i)\|_2 = 1$. Let the data distribution on the input be the uniform distribution, that is $\mathcal{D} = \text{Uniform}(\mathcal{Z})$. Consider a problem in the CLB class with a convex, 1-Lipschitz loss function $f(w, z) = -\langle w, z \rangle$, and $\mathcal{W} = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$. With this loss, the weights W_T returned by GD after T iterations are a weighted sum of the instances Z_i . As in the *birthday paradox* [MU05, Sec. 5] problem, we can show that for large d , e.g. $d = 2n^2$, the probability that any two instances from the supersample \tilde{S} share the same non-zero coordinate is smaller than some constant probability c , which is independent of the number of samples. Let E be an \tilde{S} -measurable random variable that is one if and only if no pair of instances $\tilde{Z}_{u,i}$ and $\tilde{Z}_{v,j}$ (for all $i, j \in [n]$ and all $u, v \in \{0, 1\}$) from the supersample \tilde{S} share the same coordinate.

- **Lower bound on $I(\mathcal{A}(S); U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$:** When $E = 1$ one can completely identify which instance (the index U_i) was used for training by looking at the non-zero coordinates of W_T : if $\tilde{Z}_{0,i} = \mathbf{e}(k)$ and $W_T(k) \neq 0$, then $U_i = 0$, and otherwise $U_i = 1$. Therefore, under $E = 1$, we have that $I(\mathcal{A}(S); U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) = H(U_i) = 1$.
- **Lower bound on $e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n))$:** Similarly, when $E = 1$, one can completely identify which instances (the indices U) were used for training by looking at the non-zero entries of the loss vector F : if $F_{0,i} \neq 0$, then $U_i = 0$, and otherwise $U_i = 1$. Therefore, under $E = 1$, we have that $e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n)) = H(U) = n$.

Finally, noting that this event has a constant probability, i.e. $\mathbb{P}(E = 1) \geq c$, completes the proof. □

Chapter 7

Discussion and Future Directions

This thesis explores the generalization error of learning algorithms from an information-theoretic perspective. We establish that the information complexity of a learning algorithm is a crucial factor for analyzing generalization in both distribution-dependent and minimax settings.

In Chapters 2 and 3, we provide upper bounds on the generalization error of models learned by SGLD algorithm using information complexity measures proposed in [RZ15] and [SZ20a], respectively. Our main contribution is proposing a variant of generalization bounds based on the $\text{IOMI}_{\mathcal{D}}(\mathcal{A})$ and $\text{CMI}_{\mathcal{D}}(\mathcal{A})$, which can be estimated using training samples. Our approach is effective; the bounds we obtain for SGLD are numerically non-vacuous and identify several candidates for the implicit bias of SGLD, which may be of independent interest.

In Chapter 4, we study the expressiveness of the conditional mutual information (CMI) framework of [SZ20a] and its potential for providing a unified framework for proving generalization bounds in the realizable setting. CMI is a distribution- and algorithm-dependent quantity, and it is not clear whether it is useful in distribution-dependent settings. We address this question and provide several affirmative answers in the context of binary classification. Our results show that CMI is a highly expressive measure of dependence that can be used to reason about generalization in the minimax settings.

In Chapter 5, we introduce a fundamental information complexity measure called the leave-one-out CMI (LOOeCMI), a novel information-theoretic framework for reasoning about generalization in machine learning. LOOeCMI provides upper and lower bounds on risk for 0-1 loss and interpolating learning algorithms. For consistent learners that are interpolating for any number of data, the LOOeCMI framework captures the asymptotics of risk when it converges to a non-zero quantity or to

zero polynomially. We apply the LOOeCMI framework to the one-inclusion graph algorithm and demonstrate that it yields an optimal risk bound for learning VC classes in the realizable setting. To our knowledge, no other existing information-theoretic framework can achieve optimal bounds for this setting.

Finally, in Chapter 6, we revisit the information-theoretic generalization bounds for gradient methods in the context of SCO, focusing on the GD optimization algorithm [SSSS09]. SCO with GD is arguably the simplest non-trivial test-bed to explore the limitations of the information-theoretic generalization bounds. We show that neither the CMI nor IOMI frameworks can properly characterize the excess risk of GD in SCO problems in the minmax setting. We also extend our negative results to the alternative variations of IOMI and CMI, such as evaluated CMI [SZ20a] and individual sample bounds [NHDKR19; HNKRD20; BZV20b; RBTS20; ZTL21].

7.1 Future Directions for Chapter 2 and Chapter 3

Our results in Chapter 2 and Chapter 3 show that the gradient incoherence affects the generalization performance of SGLD. However, we do not have a formal guarantee on the impact of SGLD on the gradient incoherence. It raises the following question: Does the SGLD implicitly *penalize* the gradient incoherence? A positive answer to this question can provide a complete theory of generalization for SGLD.

7.2 Future Directions for Chapter 4

1. For proper learning of VC classes, Hanneke [Han16] showed the assumption $\mathfrak{s} < \infty$ is a necessary and sufficient condition for the existence of a distribution-free bound on the expected risk of all ERMs converging at a rate $O(1/n)$. In Corollary 4.4.7 and Theorem 4.4.9, we showed the same rate for the expected risk of a broad class of ERMs can be obtained using the CMI framework. We conjecture it is possible to extend our argument to show that for a class with finite star number, every ERM has bounded eCMI.
2. An important open problem is to show that for every VC class with finite dual Helly number [BHMZ20] there exists a proper learning algorithm such that for every data distribution its expected empirical risk converges at a rate of $O(1/n)$ and it has bounded CMI. Combining the generalization guarantees that one can retrieve from Eq. (4.2) the expected excess risk of the learner with

these properties matches the optimal rate from Bousquet, Hanneke, Moran, and Zhivotovskiy [BHMZ20].

3. For improper learning of VC classes, we showed a general result for the deterministic one-inclusion graph prediction rule which is suboptimal by a $\log n$ factor. We conjecture that for every VC class with dimension d there exists a probability assignment for the randomized one-inclusion graph for which eCMI is $O(d)$. In Theorem 4.5.6, we showed this claim holds for the class of point functions.
4. In Theorem 4.5.1 we proved that eCMI is *universal* in the realizable setting. A fundamental question to ask is whether for every data distribution \mathcal{D} and consistent learner \mathcal{A} , $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}))/n$ vanishes as the number training samples n diverges at the same rate with the excess risk, i.e., $R_{\mathcal{D}}(\mathcal{A})$.

We also remark that if the answers to the above questions are affirmative, then it can be argued that the CMI framework is expressive enough so that it can explain generalization properties of VC classes. Otherwise, a negative answer to any of the questions implies that there is gap between CMI framework and VC theory.

7.3 Future Directions for Chapter 5

1. Is Theorem 5.2.5 tight? In what settings (outside those of Theorem 5.3.3) can one obtain tighter bounds? In particular, can one use the leave-one-out CMI framework to obtain tight (up to universal constants) bounds on the generalization error of arbitrary algorithms, similar to what we showed for interpolating learning algorithms?
2. Is Theorem 5.3.3 tight for finite n ? Under what conditions can we remove or tighten the binary entropy term?
3. Can one obtain an optimal bound for the one-inclusion graph algorithm under realizability via the standard CMI framework, or is there a lower bound? Is there any optimal (improper) learner for VC classes under realizability for which CMI yields optimal bounds? Lower bounds here would demonstrate that leave-one-out CMI is fundamentally stronger.
4. Leave-one-out CMI can be interpreted as an information-theoretic notion of stability. Are there connections to notions of algorithmic stability [E+03]?

In general, it is an open challenge to determine the LOO^eCMI of common learning algorithms to better understand this framework.

7.4 Future Directions for Chapter 6

Our results prompt several directions for future research:

1. One of the common properties between our constructions in Theorem 6.5.1 and Theorem 6.7.7 is that the dimension is much larger than the number of samples. In particular, we exploit the fact that the generalization guarantees of GD for SCO problems is dimension-independent in order to construct problem instances with large information complexity. In particular, it is straightforward to see that the lower bounds on IOMI and CMI that stem from our constructions in Theorem 6.5.1 and Theorem 6.7.7 depend on the dimension. It is interesting to find the minimum dimension such that there exists an SCO problem for which the information-theoretic bounds fail to characterize learnability. For the direct analysis of GD we show $\mathcal{O}(n^2)$ is sufficient (Theorem 6.7.7), while for the surrogate analysis exponential dependence, i.e., $\mathcal{O}(n^2 2^n)$ (Theorem 6.5.1), is sufficient where n is the number of training samples.
2. In this work, we proved limitations of the surrogate algorithm based on the Gaussian perturbation for the CLB subclass of SCO problems. In particular, the loss function used in Theorem 6.5.1 is a *non-smooth* convex function. It is an open question to show that such limitations exist for the subclasses of SCO problems with smooth or strongly-convex loss functions. Notice that our results in Theorem 6.7.7 suggests that a *direct* analysis still fails for the subclass of SCO problems with smooth loss functions as the loss function used for proving Theorem 6.7.7 is smooth.
3. The notable property of Gaussian perturbation is that it is instance-independent, in the sense that its structure does not depend on the problem instance, and we only need to tune the variance based on the problem instance. It is an open question to prove or refute the existence of a *instance-independent surrogate* for analyzing the generalization of gradient descent methods for SCO problems using information-theoretic frameworks. An interesting starting point is investigating the prospect of using the Gibbs algorithm [WLF16; ABTRW21] as a problem-independent surrogate.

4. We can also study the prospect of instance-dependent surrogates where the surrogate algorithm can depend on the problem instance. For this family of surrogates, the surrogate algorithm is chosen based on the data distribution, loss function, and the original learning algorithm.

A

Appendix of Chapter 2

A.1 Common Definitions

In this appendix, we collect together a few standard definitions from information theory. Let P, Q be probability measures on a common measurable space. Write $Q \ll P$ when Q is absolutely continuous with respect to P , i.e., for all measurable subsets A , $Q(A) = 0$ if $P(A) = 0$. By the Radon–Nikodym theorem, when $Q \ll P$, there exists a measurable function $\frac{dQ}{dP}$, called a Radon–Nikodym derivative or density, such that $Q(A) = \int_A \frac{dQ}{dP} dP$ for all measurable subsets A . The *KL divergence* (or *relative entropy*) of Q with respect to P , written $\text{KL}(Q \parallel P)$, is defined to be $\int \log \frac{dQ}{dP} dQ$ when $Q \ll P$ and is defined to be infinity otherwise.

Given random elements X and Y , the *mutual information between X and Y* , written $I(X; Y)$ is

$$I(X; Y) = \text{KL}(\mathbb{P}[(X, Y)] \parallel \mathbb{P}[X] \otimes \mathbb{P}[Y]),$$

where \otimes forms the product measure. Given another random element Z , the *conditional mutual information between X and Y given Z* is defined to be $I(X; Y|Z) = I(X; (Y, Z)) - I(X; Z) = I((X, Z); Y) - I(Z; Y)$.

Relative entropy and mutual information satisfy many well-known properties: For example, relative entropy and mutual information are nonnegative; $X \perp\!\!\!\perp Y \iff I(X; Y) = 0$; and $I(X; Y) \leq I(X; (Y, Z))$. From this last inequality, one may deduce that $I(X; Y) \leq I(X; Y|Z)$ when $X \perp\!\!\!\perp Z$.

A.2 Proofs of Results

A.2.1 Bounding Mutual Information by KL Divergence

The following is a well-known result that allows one to bound mutual information by the expectation of the KL divergence of a “posterior” with respect to a “prior” (where these terms are taken to have their more general interpretation from PAC-Bayesian theory, as opposed to the classical Bayesian theory).

Proposition A.2.1 (Variational Representation of Mutual Information). *Let X and Y be random elements. Then, for all probability measures P on the same space as Y ,*

$$I(X; Y) \leq \mathbb{E}[\text{KL}(\mathbb{P}^X[Y] \parallel P)],$$

with equality for $P = \mathbb{E}[\mathbb{P}^X[Y]] = \mathbb{P}[Y]$.

The result is implicit in [Kem74] and is considered folklore in the literature (e.g., it is referenced without proof in [Cat07]). For a simple derivation, see [POOAT19, Eq. (1)]. Given another random element Z , it follows immediately by the disintegration theorem [Kal06, Thm. 6.4] that, for all Z -measurable random probability measures P on the same space as Y ,

$$I^Z(X; Y) \leq \mathbb{E}^Z[\text{KL}(\mathbb{P}^{X,Z}[Y] \parallel P)] \text{ a.s.},$$

with a.s. equality for $P = \mathbb{E}^Z[\mathbb{P}^{X,Z}[Y]] = \mathbb{P}^Z[Y]$.

A.2.2 Proofs of Main Results

Proof of Theorem 2.2.3. Let \tilde{W} be a random element in \mathcal{W} such that $W \stackrel{d}{=} \tilde{W}$ and $\tilde{W} \perp\!\!\!\perp S_j^c$. Let \mathcal{G} denote the class of all functions g such that $\mathbb{E} \exp(g(\tilde{W}, S_j^c)) < \infty$. Then

$$I(W; S_j^c) = \text{KL}(\mathbb{P}(W, S_j^c) \parallel \mathbb{P}(\tilde{W}, S_j^c)) \tag{A.1}$$

$$= \sup_{g \in \mathcal{G}} \mathbb{E} g(W, S_j^c) - \log \mathbb{E} e^{g(\tilde{W}, S_j^c)} \tag{A.2}$$

where the second equality follows from the Donsker–Varadhan variational formula [BLM13, Prop. 4.15] (see also [DV75]). Let $f(w, s) = R_{\mathcal{D}}(w) - \hat{R}_s(w)$ so that $\mathbb{E} f(W, S_j^c) = \mathbb{E} R_{\mathcal{D}}(W) - \mathbb{E} \hat{R}_{S_j^c}(W)$ and $\mathbb{E} f(\tilde{W}, S_j^c) = 0$. Let ψ be the cumulant generating function of $f(\tilde{W}, S_j^c)$ and let D be the domain on which this cumulant

generating function is defined. Then $\lambda f \in \mathcal{G}$ exactly when $\lambda \in D$. Then, for every $\lambda \in D$,

$$\sup_{g \in \mathcal{G}} \mathbb{E}g(W, S_J^c) - \log \mathbb{E}e^{g(\tilde{W}, S_J^c)} \geq \lambda \mathbb{E}f(W, S_J^c) - \log \mathbb{E}e^{\lambda f(\tilde{W}, S_J^c)} \quad (\text{A.3})$$

$$= \lambda \mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W) \right] - \psi(\lambda). \quad (\text{A.4})$$

By rearranging and optimizing over λ , we find that

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W) \right] \leq \inf_{\lambda \in D} \frac{\psi(\lambda) + I(W; S_J^c)}{\lambda}.$$

Because the subset J is random and independent of (S, W) , we have $\mathbb{E}\hat{R}_{S_J^c}(W) = \mathbb{E}\hat{R}_S(W)$. Hence,

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] = \mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W) \right] \leq \inf_{\lambda \in D} \left[\frac{\psi(\lambda) + I(W; S_J^c)}{\lambda} \right].$$

At this point we have established a slightly more abstract result that permits applications beyond the subgaussian case. By the subgaussian hypothesis, $f(w, S_J^c)$ is itself σ_{n-m} -subgaussian for each $w \in \mathcal{W}$, and so the bound above reduces to

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \sqrt{2\sigma_{n-m}^2 I(W; S_J^c)}$$

using the same optimization argument as in [BZV20a], [XR17], etc. From the proof of Theorem A.3.1, $\sigma_{n-m} \leq \frac{\sigma}{\sqrt{n-m}}$, completing the proof. \square

Proof of Theorem 2.2.4. Let \tilde{W} be a random element in \mathcal{W} such that $(W, S_J, U) \stackrel{d}{=} (\tilde{W}, S_J, U)$ and $\tilde{W} \perp\!\!\!\perp S_J^c \mid \{S_J, U\}$. Let Q and P satisfy $Q(S_J, U) = \mathbb{P}^{S_J, U}[W, S_J^c]$ and $P(S_J, U) = \mathbb{P}^{S_J, U}[\tilde{W}, S_J^c]$ a.s. By the Donsker–Varadhan variational formula [BLM13, Prop. 4.15] and the disintegration theorem [Kal06, Thm. 6.4], with probability one, for all measurable functions g such that $P(S_J, U)(\exp g) < \infty$,

$$\begin{aligned} I^{S_J, U}(W; S_J^c) &= \text{KL}(Q(S_J, U) \parallel P(S_J, U)) \\ &\leq Q(S_J, U)(g) - \log P(S_J, U)(\exp g). \end{aligned}$$

Let $f(w, s) = R_{\mathcal{D}}(w) - \hat{R}_s(w)$. Note that, a.s., $P(S_J, U)(f) = \mathbb{E}^{S_J, U}[f(\tilde{W}, S_J^c)] = 0$ and

$$Q(S_J, U)(f) = \mathbb{E}^{S_J, U}[f(W, S_J^c)] = \mathbb{E}^{S_J, U}[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W)].$$

Let ψ be the cumulant generating function of $P(S_J, U)$, i.e., $\psi(\lambda; S_J, U) = \log P(S_J, U)(\exp\{\lambda f\})$. Let $D(S_J, U) = \{\lambda \in \mathbb{R} : \psi(\lambda; S_J, U) < \infty\}$. Then, with probability one, for all $\lambda \in D(S_J, U)$,

$$I^{S_J, U}(W; S_J^c) \geq \lambda \mathbb{E}^{S_J, U} \left[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W) \right] - \psi(\lambda; S_J, U).$$

Rearranging, with probability one,

$$\mathbb{E}^{S_J, U} \left[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W) \right] \leq \inf_{\lambda \in D(S_J, U)} \frac{I^{S_J, U}(W; S_J^c) + \psi(\lambda; S_J, U)}{\lambda}.$$

Because $W \perp\!\!\!\perp J$ and the subset J is random and uniformly distributed,

$$\begin{aligned} \mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] &= \mathbb{E} \mathbb{E}^{S_J, U} \left[R_{\mathcal{D}}(W) - \hat{R}_{S_J^c}(W) \right] \\ &\leq \mathbb{E} \left[\inf_{\lambda \in D(S_J, U)} \frac{I^{S_J, U}(W; S_J^c) + \psi(\lambda; S_J, U)}{\lambda} \right]. \end{aligned}$$

At this point we have established a slightly more abstract result that permits applications beyond the subgaussian case. By the subgaussian hypothesis, $f(w, S_J^c)$ is itself σ_{n-m} -subgaussian for each $w \in \mathcal{W}$, and so the bound above reduces to

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \mathbb{E} \sqrt{2\sigma_{n-m}^2 I^{S_J, U}(W; S_J^c)}$$

using the same optimization argument as in [BZV20a], [XR17], etc. From the proof of Theorem A.3.1, $\sigma_{n-m} \leq \frac{\sigma}{\sqrt{n-m}}$, completing the proof. \square

Proof of Theorem 2.2.5. For any two random measures $P(S_J, U), Q(S, U)$, the Donsker–Varadhan variational formula [BLM13, Prop. 4.15] and the disintegration theorem [Kal06, Thm. 6.4], give that with probability one

$$\text{KL}(Q(S, U) \parallel P(S_J, U)) \geq \sup_{g \in \mathcal{G}} (Q(S, U)(g) - P(S_J, U)(g) - \log [P(S_J, U)(\exp(g - P(S_J, U)(g)))]),$$

where $\mathcal{G}(S_J, U) = \{g : P(S_J, U)(\exp g) < \infty\}$.

Taking $g(w) = \lambda (R_{\mathcal{D}}(w) - \hat{R}_{S_J^c}(w))$, and letting

$$\begin{aligned} R_{\mathcal{D}}(Q) &= Q(S, U)(R_{\mathcal{D}}) & R_{\mathcal{D}}(P) &= P(S_J, U)(R_{\mathcal{D}}) \\ \hat{R}_{S_J^c}(Q) &= Q(S, U)(\hat{R}_{S_J^c}) & \hat{R}_{S_J^c}(P) &= P(S_J, U)(\hat{R}_{S_J^c}) \end{aligned}$$

where, for brevity, we have used the short hand $Q = Q(S, U)$ and $P = P(S_J, U)$. Then, with probability one

$$\begin{aligned} & \text{KL}(Q(S, U) \parallel P(S_J, U)) \\ & \geq \lambda \left(R_{\mathcal{D}}(Q) - \hat{R}_{S_J^c}(Q) - \left(R_{\mathcal{D}}(P) - \hat{R}_{S_J^c}(P) \right) \right) \\ & \quad - \log \left[P(S_J, U) \left(\exp \left(\lambda \left(R_{\mathcal{D}} - \hat{R}_{S_J^c} - \left(R_{\mathcal{D}}(P) - \hat{R}_{S_J^c}(P) \right) \right) \right) \right) \right] \end{aligned}$$

Let

$$\psi(\lambda; S, J, U) = \log \left[P(S_J, U) \left(\exp \left(\lambda \left(R_{\mathcal{D}} - \hat{R}_{S_J^c} - \left(R_{\mathcal{D}}(P) - \hat{R}_{S_J^c}(P) \right) \right) \right) \right) \right],$$

and $D(S, J, U) = \{\lambda \in \mathbb{R} : \psi(\lambda; S, J, U) < \infty\}$. With probability one

$$\left(R_{\mathcal{D}}(Q) - \hat{R}_{S_J^c}(Q) - \left(R_{\mathcal{D}}(P) - \hat{R}_{S_J^c}(P) \right) \right) \leq \inf_{\lambda \in D(S, J, U)} \frac{\text{KL}(Q(S, U) \parallel P(S_J, U)) + \psi(\lambda; S, J, U)}{\lambda}$$

Since $P(S_J, U)$ is independent of S_J^c then we have $\mathbb{E}^{S_J, J, U} \left[R_{\mathcal{D}}(P) - \hat{R}_{S_J^c}(P) \right] = 0$. Hence, by averaging over S_J^c (equivalently, taking the conditional expectation conditional on (S_J, J, U)) we have, with probability one

$$\begin{aligned} \mathbb{E}^{S_J, J, U} \left[R_{\mathcal{D}}(Q) - \hat{R}_{S_J^c}(Q) \right] &= \mathbb{E}^{S_J, J, U} \left[R_{\mathcal{D}}(Q) - \hat{R}_{S_J^c}(Q) - \left(R_{\mathcal{D}}(P) - \hat{R}_{S_J^c}(P) \right) \right] \\ &\leq \mathbb{E}^{S_J, J, U} \left[\inf_{\lambda \in D(S, J, U)} \frac{\text{KL}(Q(S, U) \parallel P(S_J, U)) + \psi(\lambda; S, J, U)}{\lambda} \right] \end{aligned}$$

Finally, by taking the full expectation, since $J \perp\!\!\!\perp Q(S, U)$ we get:

$$\mathbb{E} \left[R_{\mathcal{D}}(Q(S, U)) - \hat{R}_S(Q(S, U)) \right] \leq \mathbb{E} \left[\inf_{\lambda > 0} \frac{\text{KL}(Q(S, U) \parallel P(S_J, U)) + \psi_{S, J, U}(\lambda)}{\lambda} \right]$$

where the final $\text{KL}(Q(S, U) \parallel P(S_J, U))$ on the right hand side is between two random measures, and hence is a random variable depending on (S, J, U) ; and the expectation on the right hand side integrates over (S, J, U) .

If, for $(V \mid S_J, U) \sim P(S_J, U)$ it is the case that $\left(R_{\mathcal{D}}(V) - \hat{R}_{S_J^c}(V) \right)$ is σ -subgaussian for any (S, J, U) , then this can be optimized to get

$$\mathbb{E} \left[R_{\mathcal{D}}(Q(S, U)) - \hat{R}_S(Q(S, U)) \right] \leq \mathbb{E} \sqrt{2\sigma^2 \text{KL}(Q(S, U) \parallel P(S_J, U))}$$

When the loss is $[a_1, a_2]$ -bounded then $R_{\mathcal{D}}(V) - \hat{R}_{S_J^c}(V)$ is $\frac{a_2 - a_1}{2}$ subgaussian which

completes the proof. \square

Remark A.2.2 (Why does Theorem 2.2.5 use a boundedness assumption instead of a subgaussian assumption?). Note that we needed the boundedness assumption because even if, for $Z \sim \mathcal{D}$, $\ell(Z, w)$ was subgaussian (uniformly in $w \in \mathcal{W}$) it may not be the case that for $(V | S_J, U) \sim P(S_J, U)$, $(R_{\mathcal{D}}(V) - \hat{R}_{S_J^c}(V))$ is subgaussian. In contrast, in the proofs of Theorem 2.2.3, Theorem 2.2.4, and Theorem A.3.1 the expectations over S_j^c included in the definition of the required cumulant generating functions let us take advantage of the subgaussian property of $\ell(Z, w)$. \triangleleft

Proof of Proposition 2.2.6.

$$\text{KL}(Q_T \| P_T) \leq \text{KL}(Q_T \| P_T) + \mathbb{E}\text{KL}(Q_{|T} \| P_{|T}) = \text{KL}(Q \| P).$$

This tells us that the KL divergence between marginal distributions of the terminal parameter is upper bounded by the KL between the distributions of the full trajectories.

Assuming $Q_0 = P_0$, we may decompose $\text{KL}(Q \| P)$ across iterations, obtaining

$$\text{KL}(Q \| P) = \mathbb{E}_{W \sim Q} \left[\log \frac{dQ}{dP}(W) \right] = \mathbb{E}_{W \sim Q} \left[\sum_{t=1}^T \log \frac{dQ_{t|}}{dP_{t|}}(W) \right] = \sum_{t=1}^T \mathbb{E}_{Q_{0:(t-1)}} [\text{KL}(Q_{t|} \| P_{t|})]. \quad (\text{A.5})$$

\square

A.3 Mutual Information Bound for Subgaussian Losses

Theorem A.3.1 (Xu and Raginsky's Theorem 1). *Suppose that $\ell(w, Z)$ is σ -subgaussian when $Z \sim \mu$, for all $w \in \mathcal{W}$, Then*

$$|\mathbb{E} [R_{\mathcal{D}}(W) - \hat{R}_S(W)]| \leq \sqrt{\frac{2\sigma^2}{n} I(S; W)}$$

A proof of this result is found in [XR17]. However, one may use the arguments therein to establish the further conclusion that $\ell(W, Z)$ or $R_S(W)$ is also subgaussian, which is not generally true. In this section we briefly describe the flaw in that logic and provide a clarification of their proof under the same assumptions. [RZ15] give

a proof for discrete parameter spaces, which does not contain this flaw. While it is straightforward to cast their proof into measure-theoretic language, we give the details for completeness.

The discussion in [XR17] preceding the theorem asserts that if $f : \mathcal{W} \times \mathcal{S}$ is such that $f(w, S)$ is σ subgaussian for all $w \in \mathcal{W}$ and if $W \perp\!\!\!\perp S$ then $f(W, S)$ is σ -subgaussian. A simple counter example is given by $\mathcal{W} = \mathcal{S} = \mathbb{R}$, with $f(w, s) = w + s$, and $(W, S) \sim \text{Cauchy} \times N(0, 1)$. In this case $f(w, S)$ is clearly 1-subgaussian for each $w \in \mathcal{W}$, while $f(W, S)$ does not even have bounded absolute first moment, let alone a moment generating function defined in any open ball about 0.

The main issue in the argument establishing subgaussianity of $f(W, S)$ is failing to properly use a version of the conditional variance formula (modified to apply for moment generating functions as opposed to variances). The intuition of the conditional variance formula is useful in reconciling the final result with our counterexample, but is not sufficient for a general proof as the subgaussian parameter is not generally a standard deviation. The conditional variance formula asserts that

$$\text{Var}(f(W, S)) = \mathbb{E} [\text{Var}^W f(W, S)] + \text{Var} (\mathbb{E}^W f(W, S)).$$

The argument by which one would conclude that $f(W, S)$ is subgaussian only acknowledges the first term, thus assuming that the second term is 0 (which would only hold when $\mathbb{E}^W f(W, S)$ is a.s. constant in W).

More precisely, since we are working with subgaussian parameters instead of true standard deviations:

$$\begin{aligned} & \log \mathbb{E} \exp(t(f(W, S) - \mathbb{E}f(W, S))) \\ &= \log \mathbb{E} [\exp(t(\mathbb{E}^W f(W, S) - \mathbb{E}f(W, S))) \mathbb{E}^W \exp(t(f(W, S) - \mathbb{E}^W f(W, S)))] \\ &\leq \log \exp(t^2 \sigma^2 / 2) \mathbb{E}[\exp(t(\mathbb{E}^W f(W, S) - \mathbb{E}f(W, S)))] \\ &= t^2 \sigma^2 / 2 + \log \mathbb{E}[\exp(t(\mathbb{E}^W f(W, S) - \mathbb{E}f(W, S)))] \end{aligned}$$

The RHS is $\geq t^2 \sigma^2 / 2$ with equality if and only if $(\mathbb{E}^W f(W, S) - \mathbb{E}f(W, S))$ is constant (by Jensen' inequality). The first inequality is an equality when $f(w, S)$ is normal with variance σ^2 for all $w \in \mathcal{W}$.

Ergo, the assertion that $f(W, S)$ is σ -subgaussian holds exactly when $(\mathbb{E}_S f(W, S) - \mathbb{E}f(W, S))$ is constant. This situation is not generally of interest in learning theory; this amounts to saying that all parameter vectors lead to the same expected generalization error, and hence there is no purpose to learning from the data!

The final result is, of course, still valid and may be proven directly via the Donsker–Varadhan variational formula.

Proof. As in [XR17] we will leverage the fact that for each $w \in \mathcal{W}$, $f(w, S) = \frac{1}{n} \sum_{i=1}^n \ell(w, Z_i)$ is $\tau = \sigma/\sqrt{n}$ subgaussian, *however these variable may have different means for each value of w .* Let $\check{f}(w, s) = f(w, s) - \mathbb{E}f(w, S)$.

By Donsker–Varadhan and the fact that $\mathbb{E}^W \check{f}(\bar{W}, \bar{S}) = 0$ a.s.,

$$\begin{aligned} I(W; S) &\geq \mathbb{E} \lambda \check{f}(W, S) - \log \mathbb{E} \exp(\lambda \check{f}(\bar{W}, \bar{S})) \\ &\geq \mathbb{E} \lambda \check{f}(W, S) - \log \mathbb{E} \mathbb{E}^W \exp(\lambda \check{f}(\bar{W}, \bar{S})) \\ &\geq \lambda \mathbb{E} \check{f}(W, S) - \log \mathbb{E} \exp(\lambda^2 \tau^2 / 2) \\ &\geq \lambda \mathbb{E} \check{f}(W, S) - \lambda^2 \tau^2 / 2. \end{aligned}$$

Optimizing over λ now yields the desired result, because

$$|\mathbb{E} \check{f}(W, S)| = |\mathbb{E} [f(W, S) - \mathbb{E}^W f(W, \bar{S})]| = |\mathbb{E} [R_D(W) - \hat{R}_S(W)]|.$$

□

A.4 Properties of the Hypergeometric Distribution and of Finite Population Variances

In this section, we enumerate a number of well-known results, and also derive some particular ones for our application.

A.4.1 Properties of the Hypergeometric Distribution

Let $n, m, b \in \mathbb{N}$, $m, b \leq n$. Write $B \sim \text{HG}(n, m, b)$ when

$$\mathbb{P}(B = j) = \frac{\binom{m}{j} \binom{n-m}{b-j}}{\binom{n}{b}}, \quad j \in \{0 \vee b + m - n, \dots, n \wedge m\}.$$

It follows that

$$\mathbb{E}(B) = b \frac{m}{n} \text{Var}(B) = b \frac{m}{n} \frac{n-m}{n} \frac{n-b}{n-1} \leq b \frac{m(n-m)}{n^2}$$

A.4.2 Finite Population Statistics with Disjoint Samples

In this section we compute the covariance of the sample means for each population, and provide a formula for the variance of a linear combination of the two estimators.

Lemma A.4.1 (Variance for disjoint finite population statistics). *Suppose that there is a finite population of size, N , $S = (y_1, \dots, y_N)$. Consider two disjoint subsets of sizes n_1 and n_2 are chosen uniformly at random from S . Let \bar{Y}_i be the sample mean on the i th sample. Let Σ be the population variance matrix. Then*

$$\text{Var} \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{pmatrix} = \frac{1}{N-1} \begin{bmatrix} (N-n_1)/n_1 & -1 \\ -1 & (N-n_2)/n_2 \end{bmatrix} \otimes \Sigma$$

$$\text{Var}(a\bar{Y}_1 - b\bar{Y}_2) = \frac{1}{(N-1)} \left(-(a-b)^2 + N(a^2/n_1 + b^2/n_2) \right) \Sigma$$

Proof. Let ζ_i be an indicator for whether y_i appears in the first sample, and let W_i be an indicator for whether y_i appears in the second sample.

Let $\mu = \frac{1}{N} \sum_{i=1}^N y_i$ and let $\Sigma = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)(y_i - \mu)'$

Then for any $i \neq j$:

$$\begin{aligned} \zeta_i &\sim \text{Ber}(n_1/N) & W_i &\sim \text{Ber}(n_2/N) \\ \text{Var}(\zeta_i) &= \frac{n_1(N-n_1)}{N^2} & \text{Var}(W_i) &= \frac{n_2(N-n_2)}{N^2} \end{aligned}$$

$$\begin{aligned} \text{Cov}(\zeta_i, \zeta_j) &= \mathbb{E}[\zeta_i \zeta_j] - \frac{n_1^2}{N^2} & \text{Cov}(W_i, W_j) &= \mathbb{E}[W_i W_j] - \frac{n_2^2}{N^2} \\ &= \mathbb{P}[\zeta_i = \zeta_j = 1] - \frac{n_1^2}{N^2} & &= \text{PP}[W_i = W_j = 1] - \frac{n_2^2}{N^2} \\ &= \frac{n_1(n_1-1)}{N(N-1)} - \frac{n_1^2}{N^2} & &= \frac{n_2(n_2-1)}{N(N-1)} - \frac{n_2^2}{N^2} \\ &= -\frac{n_1}{N} \left(1 - \frac{n_1}{N}\right) \frac{1}{N-1} & &= -\frac{n_2}{N} \left(1 - \frac{n_2}{N}\right) \frac{1}{N-1} \end{aligned}$$

$$\begin{aligned}
\text{Cov}(\zeta_i, W_i) &= \mathbb{E}[\zeta_i W_i] - \frac{n_1 n_2}{N^2} \\
&= \mathbb{P}[\zeta_i = W_j = 1] - \frac{n_1 n_2}{N^2} \\
&= 0 - \frac{n_1 n_2}{N^2} \\
&= -\frac{n_1 n_2}{N^2}
\end{aligned}
\qquad
\begin{aligned}
\text{Cov}(\zeta_i, W_j) &= \mathbb{E}[\zeta_i W_j] - \frac{n_1 n_2}{N^2} \\
&= \text{PP}[\zeta_i = W_j = 1] - \frac{n_1 n_2}{N^2} \\
&= \frac{n_1 n_2}{N(N-1)} - \frac{n_1 n_2}{N^2} \\
&= \frac{n_1 n_2}{N^2(N-1)}.
\end{aligned}$$

$$(\bar{Y}_1, \bar{Y}_2) = \sum_{i=1}^N y_i (\zeta_i/n_1, W_i/n_2)$$

$$\begin{aligned}
\text{Var}(\bar{Y}_1) &= \text{Var} \left(\sum_{i=1}^N \frac{y_i}{n_1} \zeta_i \right) \\
&= \frac{1}{n_1^2} \left(\sum_{i=1}^N y_i y_i' \frac{n_1(N-n_1)}{N^2} - \sum_{i \neq j} y_i y_j' \frac{n_1(N-n_1)}{N^2(N-1)} \right) \\
&= \frac{(N-n_1)}{n_1 N^2} \left(\sum_{i=1}^N y_i y_i' - \sum_{i \neq j} y_i y_j' \frac{1}{N-1} \right) \\
&= \frac{(N-n_1)}{n_1(N-1)N} \sum_{i=1}^N (y_i - \mu)(y_i - \mu)' \\
&= \frac{(N-n_1)}{n_1(N-1)} \Sigma
\end{aligned}$$

Similarly

$$\text{Var}(\bar{Y}_2) = \frac{(N-n_2)}{n_2(N-1)} \Sigma$$

Now, for the less well known part:

$$\begin{aligned}
\text{Cov}(\bar{Y}_1, \bar{Y}_2) &= \text{Cov}\left(\sum_{i=1}^N \frac{y_i}{n_1} \zeta_i, \sum_{i=1}^N \frac{y_i}{n_2} W_i\right) \\
&= \sum_{i=1}^N \frac{y_i y'_i}{n_1 n_2} \text{Cov}(\zeta_i, W_i) + \sum_{i \neq j} \frac{y_i y'_j}{n_1 n_2} \text{Cov}(\zeta_i, W_j) \\
&= -\sum_{i=1}^N \frac{y_i y'_i}{n_1 n_2} \frac{n_1 n_2}{N^2} + \sum_{i \neq j} \frac{y_i y'_j}{n_1 n_2} \frac{n_1 n_2}{N^2(N-1)} \\
&= -\frac{1}{N^2} \left(\sum_{i=1}^N y_i y'_i - \sum_{i \neq j} y_i y'_j \frac{1}{N-1} \right) \\
&= -\frac{1}{N-1} \Sigma
\end{aligned}$$

Hence

$$\text{Var} \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{pmatrix} = \frac{1}{N-1} \begin{bmatrix} (N-n_1)/n_1 & -1 \\ -1 & (N-n_2)/n_2 \end{bmatrix} \otimes \Sigma$$

For our application we need $\text{Var}(a\bar{Y}_1 - b\bar{Y}_2)$:

$$\begin{aligned}
\text{Var}(a\bar{Y}_1 - b\bar{Y}_2) &= a^2 \frac{(N-n_1)}{n_1(N-1)} \Sigma + b^2 \frac{(N-n_2)}{n_2(N-1)} \Sigma + 2ab \frac{1}{N-1} \Sigma \\
&= \frac{1}{(N-1)} \left(a^2 \frac{N-n_1}{n_1} + 2ab + b^2 \frac{N-n_2}{n_2} \right) \Sigma \\
&= \frac{1}{(N-1)} \left(-(a-b)^2 + N(a^2/n_1 + b^2/n_2) \right) \Sigma
\end{aligned}$$

□

Lemma A.4.2 (Bounding $\mathbb{E}\mathbb{E}^{S_J, J, U} \|\xi_t\|_2^2$ for SGLD). *In the setting of Section 2.3.1*

$$\mathbb{E}\mathbb{E}^{S_J, J, U} \|\xi_t\|_2^2 = \frac{n(n-m)}{(n-1)^2 b_t} \left(1 + \frac{b_t n - m - 1}{n} \frac{1}{m} \right) \mathbb{E}[\hat{\Sigma}_t(S)]$$

Proof. Applying the conditional variance formula gives:

$$\begin{aligned}
\mathbb{E}\mathbb{E}^{S_J, J, U} \|\xi_t\|_2^2 &= \mathbb{E}\text{Var}^{S, W_t}(\mathbb{E}^{b_t, W_t, S}[\xi_t]) + \mathbb{E}\mathbb{E}^{S, W_t}[\text{Var}^{b_t, W_t, S}(\xi_t)] \\
&= 0 + \mathbb{E}\mathbb{E}^{S, W_t} \left[\text{Var}^{b_t, W_t, S} \left(\frac{b_t^c}{b_t} \nabla \tilde{R}_{S_t^c}(W_t) - \frac{b_t^c}{b_t} \nabla \tilde{R}_{S_J}(W_t) \right) \right] \\
&= \mathbb{E}\mathbb{E}^{S, W_t} \left[\frac{(b_t^c)^2}{b_t^2} \text{Var}^{b_t, W_t, S} \left(\nabla \tilde{R}_{S_t^c}(W_t) - \nabla \tilde{R}_{S_J}(W_t) \right) \right]
\end{aligned}$$

Applying Lemma A.4.1 further yields

$$\begin{aligned}
&\mathbb{E}^{S, W_t} \left[\frac{(b_t^c)^2}{b_t^2} \text{Var}^{b_t, W_t} \left(\nabla \tilde{R}_{S_t^c}(W_t) - \nabla \tilde{R}_{S_J}(W_t) \right) \right] \\
&= \mathbb{E}^{S, W_t} \left[\frac{(b_t^c)^2}{b_t^2} \frac{1}{(n-1)} \left(\frac{n}{b_t^c} + \frac{n}{m} \right) \hat{\Sigma}_t(S) \right] \\
&= \frac{n}{(n-1)b_t^2} \mathbb{E}^{S, W_t} \left[b_t^c + (b_t^c)^2 \frac{1}{m} \right] \mathbb{E}^{S, W_t}[\hat{\Sigma}_t(S)] \\
&= \frac{n}{(n-1)b_t^2} \left(b_t \frac{n-m}{n} + \left(\frac{(n-m)^2}{n^2} b_t^2 + b_t \frac{m}{n} \frac{n-m}{n} \frac{n-b_t}{n-1} \right) \frac{1}{m} \right) \mathbb{E}^{S, W_t}[\hat{\Sigma}_t(S)] \\
&= \frac{n}{(n-1)b_t^2} \left(b_t \frac{n-m}{n-1} + b_t^2 \frac{(n-m)(n-m-1)}{n(n-1)m} \right) \mathbb{E}^{S, W_t}[\hat{\Sigma}_t(S)] \\
&= \frac{n}{(n-1)b_t^2} \left(b_t \frac{n-m}{n-1} + b_t^2 \frac{(n-m)(n-m-1)}{n(n-1)m} \right) \mathbb{E}^{S, W_t}[\hat{\Sigma}_t(S)] \\
&= \frac{n(n-m)}{(n-1)^2 b_t} \left(1 + \frac{b_t}{n} \frac{n-m-1}{m} \right) \mathbb{E}^{S, W_t}[\hat{\Sigma}_t(S)]
\end{aligned}$$

□

A.5 Asymptotic Results

A.5.1 Langevin Dynamics

In this section we continue from the end of Section 2.3.2. Under the assumption that $\tilde{\ell}$ is L -Lispchitz (the same assumption as in [BZV20a]) we have the following results which portray the asymptotic behavior of the expected generalization error of the Langevin diffusion algorithm for ℓ being the 0-1 loss (which is 1/2-subgaussian):

$$\mathbb{E}_{W_T \sim Q_T} (R_{\mathcal{D}}(W_T) - R_S(W_T)) \leq \frac{L}{2(n-1)} \sqrt{\sum_{t=1}^T \beta_t \eta_t}$$

Geometrically Decaying Learning Rate

Under an assumption of L -Lipschitz loss and geometrically decaying learning rate and a temperature that ramps up to a polynomial in n ($\eta_t = \eta_0 \rho^t$ for $0 < \rho < 1$ and that $\beta_t = \beta_0(n-1)^\theta(1-\nu^t)$ for some $0 < \theta < 1$) then we have the following bound:

$$\sup_{T \geq 0} \left[\mathbb{E}_{W_T \sim Q_T} (R_{\mathcal{D}}(W_T) - R_S(W_T)) \right] \leq \frac{L}{2(n-1)^{1-\theta}} \sqrt{\beta_0 \eta_0 \frac{\rho(1-\nu)}{(1-\rho)(1-\rho\nu)}}$$

Polynomial Decaying Learning Rate

Under an assumption of L -Lipschitz loss and polynomial decaying learning rate and temperature that is polynomial in n ($\eta_t = \eta_0 t^{-\alpha}$ for $\alpha > 0$ and that $\beta_t = \beta_0(n-1)^p$ for some $0 < p < 1$) then we have the following bound:

$$\left[\mathbb{E}_{W_T \sim Q_T} (R_{\mathcal{D}}(W_T) - R_S(W_T)) \right] \leq \begin{cases} \frac{L}{2(n-1)^{1-p}} \sqrt{1 + \frac{1}{\alpha-1} T^{1-\alpha}} & \alpha < 1 \\ \frac{L}{2(n-1)^{1-p}} \sqrt{1 + \log(T)} & \alpha = 1 \\ \frac{L\alpha}{2(n-1)^{1-p(\alpha-1)}} & \alpha > 1 \end{cases}$$

A.6 Comparing Theorems 2.2.3 to 2.2.5 when $m = n - 1$

Let $V \sim P(S_J, U)$, $W \sim Q(S, U)$, and $\tilde{W} \sim Q(S, U)$ independently of W . In the case of $[a_1, a_2]$ -bounded loss, $(R_{\mathcal{D}}(V) - \hat{R}_{S \setminus S}(V))$ is $(a_2 - a_1)/2$ -subgaussian, so that Theorem 2.2.5 yields:

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(\tilde{W}) \right] \leq \mathbb{E} \sqrt{(a_2 - a_1)^2 \text{KL}(Q(S, U) \| P(S_J, U)) / 2}.$$

Using KL divergence based upper bounds for mutual information (Proposition A.2.1), Theorem 2.2.4 gives us

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \mathbb{E} \sqrt{(a_2 - a_1)^2 \mathbb{E}^{S_J, U} [\text{KL}(Q(S, U) \| P(S_J, U))]} / 2,$$

while Theorem 2.2.3 yields:

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \leq \sqrt{(a_2 - a_1)^2 \mathbb{E} [\text{KL}(Q(S, U) \| P(S_J, U))]} / 2$$

for $m = n - 1$, the bounds are ranked as [2.2.3](#) \geq [2.2.4](#) \geq [2.2.5](#) (by Jensen's inequality for each conditional expectation being passed into $\sqrt{\cdot}$). When $\text{KL}(Q(S) \| P(S_J))$ has a large variance then the difference can be quite material.

A.7 An analytically tractable example

We present a simple analytic example, where our upper bound is a clear improvement over existing work when similar simplifications are performed. Let $S = \{z_1, \dots, z_n\} \sim \mathcal{D}^n$ be a sample from the distribution \mathcal{D} on \mathbb{R} . We wish to estimate the mean of \mathcal{D} , μ . We will use the loss function $\ell(z, w) = \tilde{\ell}(z, w) = (z - w)^2$ where $w \in \mathcal{W} = \mathbb{R}$. The distribution \mathcal{D} , is assumed to satisfy the sub-Gaussianity assumption in [Theorems 2.2.3](#) and [2.2.5](#) for this loss. Upon specializing the SGLD update rule [\(2.7\)](#) to this setting:

$$W_{t+1} = W_t - \eta_t \frac{d}{dW_t} \tilde{R}_S(W_t) + \sqrt{\frac{2\eta_t}{\beta}} \epsilon_t = \left(1 - \frac{2\eta_t}{n}\right) W_t + \frac{2\eta_t}{n} \sum_{i=1}^n z_i + \sqrt{\frac{2\eta_t}{\beta}} \epsilon_t. \quad (\text{A.6})$$

We will apply the data-dependent generalization bound in [Theorem 2.2.5](#) with $m = |S_J| = n - 1$ and set $\{i^*\} = J$. Since we are working with LD, we set the random variable U to a constant (trivial random variable). It follows that:

$$\text{KL}(Q_{t+1|1}(S) \| P_{t+1|1}(S_J)) = \frac{(\mu_{t+1} - \mu'_{t+1})^2}{4\eta_t/\beta} = \frac{\beta}{n^2} z_{i^*}^2 \eta_t. \quad (\text{A.7})$$

Thus the expected generalization error is upper bounded by:

$$\mathbb{E} \sqrt{2\sigma^2 \text{KL}(Q_T(S) \| P_T(S_J))} \leq \mathbb{E} \sqrt{2\sigma^2 \frac{\beta}{n^2} z_{i^*}^2 \sum_{t=0}^{T-1} \eta_t} = \mathbb{E}[|z_i|] \left(\sqrt{2\sigma^2 \frac{\beta}{n^2} \sum_{t=0}^{T-1} \eta_t} \right). \quad (\text{A.8})$$

When one applies the results in [\[XR17; PJJ18\]](#), the upper bounded on the generalization error can be shown to be:

$$\sqrt{\frac{2\sigma^2}{n} I(W_T; S)} \leq \sqrt{\frac{2\sigma^2}{n} \sum_{t=0}^{T-1} I(\bar{W}_{t+1}; S | W_1^t)} \leq \sqrt{2\sigma^2 \frac{\beta}{n^2} E[z_i^2] \sum_{t=0}^{T-1} \eta_t} \quad (\text{A.9})$$

Comparing with [\(A.8\)](#) we see that this bound can be is larger since $E[|z_i|] \leq \sqrt{E[z_i^2]}$ from Jensen's inequality. The discrepancy can be made arbitrarily large based on the

choice of \mathcal{D} .

A.8 Experiment Details

The first architecture and dataset we consider is a three-layer multilayer perceptron (MLP), with 600 hidden units per hidden layer and rectified linear unit (ReLU) activation functions, trained on MNIST [LCB10]. In Fig. 2.1a, we compare the bound for two amounts of held out data, $n - m = |S_j^c|$. We see that the empirical performance reflects our analytical results that the bound is tighter for large m . As can be inferred from Eq. (2.4), the difference between $\|\xi_t\|^2$ and $\|\nabla \tilde{R}_t\|^2$ increases with m .

The remainder of our experiments consider convolutional neural networks (CNNs). For MNIST and Fashion-MNIST, we use a standard network configuration with two convolutional layers (with 32 and 64 filters of size 5×5 , respectively, followed by 2×2 max pooling after each convolutional layer), followed by two fully connected layers (1024 nodes each) with ReLU activations.

Our final experiment uses the CIFAR-10 dataset. The CNN architecture has two convolutional layers and three fully connected layers. Both convolutional layers use 64 filters of size 5×5 . After each convolutional layer there is a 2×2 max pooling layer. Then, we have three fully connected layers with the number of neurons 384, 192, and 10 respectively.

A.8.1 Evaluation of the generalization bound

We estimate our generalization error bound and that of [MWZZ18] using nested Monte Carlo simulations. We use the results of Theorem 2.3.1, specifically Eq. (2.6). In order to evaluate this bound we perform two Monte Carlo estimate: one for $\mathbb{E}^{S,J,U} \|\xi_t\|_2^2$, and then for the full expectation (outside of the $\sqrt{\cdot}$). For our bound, for each hyperparameter combination, we have used 10 simulations for the outer expectation, each using 10 simulations to estimate the inner expectation. For the generalization bound in Mou, L. Wang, Zhai, and Zheng [MWZZ18] we have used their 100 simulations to evaluate the bound given by their Theorem 10.

A.8.2 Learning Rate and Inverse Temperatures for Figs. 2.1b and 2.1c

In Fig. 2.1b, we use

$$\beta_t^{(\text{high})} = 100 \times \max\left\{\exp\left(\frac{t}{100}\right), 55000\right\} \quad (\text{A.10})$$

$$\beta_t^{(\text{low})} = 100 \times \max\left\{\exp\left(\frac{t}{100}\right), 5000\right\} \quad (\text{A.11})$$

where t denotes the iteration number.

All other parameters are the same and are outlined in Table A.2.

For the ‘high’ inverse temperature schedule, at Iteration 5 the training error is 3.21% and the generalization error is 0.9%, while for the ‘low’ inverse temperature schedule, at epoch 5 the training error is 5.18% and the generalization error is 0.16%.

In Fig. 2.1c we consider $\eta_t^{(\text{small})} = 8 \times 10^{-4} \times 0.96^{\left(\frac{t}{2000}\right)}$ and $\eta_t^{(\text{large})} = 2 \times 10^{-3} \times 0.96^{\left(\frac{t}{2000}\right)}$, and the rest of the parameters are the same and are outlined in Table A.2. For the ‘small’ learning rate, the training error and the test-set generalization error at Epoch 6 for the small learning rate scenario are 7.62% and 1.1% , respectively;; while for the ‘large’ learning rate the training error and the test-set generalization error at Epoch 6 are 6.3% and 1.0%, respectively.

A.8.3 Hyperparameters of our experiments

In Tables A.1 to A.4, we provide the hyperparameter and training details of the experiments that were presented in Section 2.4.

Parameter	Values
Dataset	MNIST
Architecture	MLP with 3 hidden layers
Batch size	100
Learning rate	learning rate= 8×10^{-3} , decay steps=600, decay rate=0.95
Beta schedule	$\min\{10 \times \exp(\text{iter}/400), 2000\}$
Number of epochs	15
Average Final training error	1.40%
Average Final test error	4.12%
# training examples	55000
Number of runs	50

Table A.1: Details of Experiments reported in Fig. 2.1a for MNIST with MLP

Parameter	Values
Dataset	MNIST
Architecture	CNN with 2 conv. layers
Batch size	100
Learning rate	learning rate= 4×10^{-3} , decay steps=2000, decay rate=0.96
Beta schedule	$\min\{10 \times \exp(\text{iter}/100), 55000\}$
Number of epochs	15
Average Final training error	1.81%
Average Final test error	2.03%
# training examples	55000
Number of runs	50

Table A.2: Details of Experiments reported in Figs. 2.1b to 2.1d for MNIST with CNN

Parameter	Values
Dataset	Fashion-MNIST
Architecture	CNN with 2 conv. layers
Batch size	100
Learning rate	learning rate= 4×10^{-3} , decay steps=3500, decay rate=0.93
Beta schedule	$\min\{10 \times \exp(\text{iter}/100), 55000\}$
Number of epochs	25
Average Final training error	8.3%
Average Final test error	10.83%
# training examples	60000
Number of runs	20

Table A.3: Details of Experiments reported in Fig. 2.1e for Fashion-MNIST

A.9 High Probability PAC-Bayes Bounds

We can leverage the methods used to provide bounds on the expected generalization error above to also derive high probability bounds for the generalization error. We will give an example of this here for completeness, though more work can be done to select a tighter bound from more recent literature and to tune the parameters available to optimize the bound further. For example, in our setting we could optimally tune the level of data dependence for the bound to be tightened. We will make use of Shalev-Shwartz and Ben-David [SB14] (theorem 31.1 therein), which we state here under the notation and definitions of our work, and in the context of Section 2.3.1.

Proposition A.9.1 ([SB14] Theorem 31.1). *Suppose that the loss function is $[0, 1]$ -bounded. Let P be any prior distribution. With probability at least $(1 - \delta)$ (over the choice of $S \sim \mathcal{D}^n$) for any posterior distribution Q (even those depending on S) with*

Parameter	Values
Dataset	CIFAR-10
Architecture	CNN with 2 conv. layers
Batch size	200
Learning rate	learning rate= 5×10^{-3} , decay steps=2000, decay rate=0.95
Beta schedule	$\min\{10 \times \exp(\text{iter}/100), 55000\}$
Number of epochs	50
Average Final training error	6.9%
Average Final test error	29.9%
$ S_J $	$\text{len}(\text{training_set})-1$
# training examples	50000
Number of runs	30

Table A.4: Details of Experiments reported in Fig. 2.1f for CIFAR-10

$W \sim Q$,

$$\mathbb{E}^S \left[R_{\mathcal{D}}(W_T) - \hat{R}_S(W_T) \right] \leq \sqrt{\frac{\text{KL}(Q \| P) + \log(n/\delta)}{2(n-1)}}$$

In our setting P will be allowed to depend on m data points chosen uniformly at random, while Q will depend on the full dataset, so we can apply this result conditional on the subset upon which P depends. Therefore, for any $S_J \in \mathcal{Z}^m$ and any $U \in \mathcal{U}$ and for any kernel $P : \mathcal{Z}^n \times \mathcal{U} \rightarrow \mathcal{M}_1(\mathcal{W}^T)$ be any prior distribution which depends on S_J , with probability at least $(1 - \delta)$ (over the choice of $S_J^c \sim \mathcal{D}^{n-m}$) for any posterior distribution Q (even those depending on S_J^c) with $W \sim Q$,

$$\begin{aligned} \mathbb{E}^S \left[R_{\mathcal{D}}(W_T) - \hat{R}_{S_J^c}(W_T) \right] &\leq \sqrt{\frac{\text{KL}(Q(S, U) \| P(S_J, U)) + \log((n-m)/\delta)}{2(n-m-1)}} \\ &\leq \sqrt{\frac{\sum_{t=1}^T \frac{\beta_t \eta_t}{8} \mathbb{E}^{S, J, U} \|\xi_t\|_2^2 + \log((n-m)/\delta)}{2(n-m-1)}} \end{aligned}$$

In the case of Langevin dynamics when using worst case, Lipschitz constant base upper bounds, this gives

$$\mathbb{E}^S \left[R_{\mathcal{D}}(W_T) - \hat{R}_{S_J^c}(W_T) \right] \leq \sqrt{\frac{\frac{L^2}{(n-1)(m-1)} \sum_{t=1}^T \frac{\beta_t \eta_t}{8} + \log((n-m)/\delta)}{2(n-m-1)}}$$

which provides a less trivial tradeoff between m and $n - m$ compared to the expected generalization error bound. One could further take expectations over U and/or J to

get high probability bounds for the generalization error based on the full empirical loss.

We intend to investigate such bounds further in future work, and this section serves merely to illustrate the possibility and nature of such high-probability bounds based on data-dependent estimates of mutual information and data-dependent PAC-Bayes priors. We acknowledge that these are not the tightest such bounds possible.

B

Appendix of Chapter 3

B.1 CMI, Membership Attack, and Fano's Inequality

Let $\tilde{Z}k$, $U^{(k)}$, and S as in Definition 3.1.2. Consider the following hypothesis testing problem. Assume a decision maker observes W and wishes to recover $U^{(k)}$ by having access to the super-sample $\tilde{Z}k$. For any estimate $\hat{U} = \Psi(W, \tilde{Z}k)$, we have the Markov chain

$$U^{(k)} \rightarrow S \rightarrow W \rightarrow \widehat{U^{(k)}}.$$

and so, combined with the fact that $U^{(k)}$ is uniformly distributed over a set of size k^n , we can invoke Fano's inequality to bound the error probability of the decision maker. In particular,

$$\inf_{\Psi} \mathbb{P} \left[\Psi \left(W, \tilde{Z}k \right) \neq U^{(k)} \right] \geq 1 - \frac{I(W; U^{(k)} | \tilde{Z}k) + \log 2}{n \log k}.$$

Hence, $I(W; U^{(k)} | \tilde{Z}k)$ provides a lower bound on the hardness of the hypothesis testing problem, where one wants to identify the training sample given access to $\tilde{Z}k$ and W .

Some interpretation of our result is helpful. Consider an adversary who has access to the supersample $\tilde{Z}k$ and wishes to identify the training set that was used for the training after observing the output of a learning algorithm W . Our result here showed that the CMI upperbounds the success probability of *every* adversary. Also, recall that the CMI upper bounds the expected generalization error. In the literature of data privacy in machine learning, this problem is known as Membership Attack [SSSS17],

and it is empirically observed that a machine learning model leaks information about its training set when the generalization error is large [SSSS17]. Our result in this section provides a formal connection between generalization and this specific membership attack problem.

B.2 Matching the leading coefficient of Theorem 3.1.1 with $\text{CMI}_{\mathcal{D}}(\mathcal{A})$

Theorem B.2.1. *Let $\text{CMI}_{\mathcal{D}}(\mathcal{A})$ as defined in the introduction. Then, for $k > 2$*

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{\text{CMI}_{\mathcal{D}}(\mathcal{A})}{\lambda^*} + \frac{\exp(\frac{\lambda^*}{n}) - \frac{\lambda^*}{n} - 1}{\frac{\lambda^*}{n}} \frac{k^3 + 7k^2 - 8k - 16}{4(k^3 - 2k^2)},$$

where

$$\lambda^* = n\mathfrak{W}\left(\left(\frac{4(k^3 - 2k^2)\text{CMI}_{\mathcal{D}}(\mathcal{A})}{n(k^3 + 7k^2 - 8k - 16)} - 1\right) \exp(-1)\right) + n,$$

and \mathfrak{W} is the Lambert W function.

The proof is deferred to Appendix B.3. Here, we quantitatively compare Theorem 3.1.1, Theorem 3.1.3, and Theorem B.2.1, we consider the case that the output of \mathcal{A} takes value in a finite set and $k \rightarrow \infty$. In this case, Theorem B.2.1 can be rephrased as

$$\begin{aligned} \text{EGE}_{\mathcal{D}}(\mathcal{A}_n) &\leq \lim_{k \rightarrow \infty} \frac{\text{CMI}_{\mathcal{D}}(\mathcal{A})}{\lambda^*} + \frac{\exp(\frac{\lambda^*}{n}) - \frac{\lambda^*}{n} - 1}{\frac{\lambda^*}{n}} \frac{k^3 + 7k^2 - 8k - 16}{4(k^3 - 2k^2)} \\ &= \frac{\text{IOMI}_{\mathcal{D}}(\mathcal{A})}{\lambda_{\infty}^*} + \frac{\exp(\frac{\lambda_{\infty}^*}{n}) - \frac{\lambda_{\infty}^*}{n} - 1}{4\frac{\lambda_{\infty}^*}{n}} \end{aligned} \quad (\text{B.1})$$

where $\lambda_{\infty}^* = n\mathfrak{W}\left(\left(\frac{4\text{IOMI}_{\mathcal{D}}(\mathcal{A})}{n} - 1\right) \exp(-1)\right) + n$. In the next plot, we compare the values of the bounds in Theorem 3.1.1, Theorem 3.1.3, and Theorem B.2.1 assuming $\text{IOMI}_{\mathcal{D}}(\mathcal{A}) = 1$. As seen, the bound in Theorem B.2.1 provides much tighter constant compared with Theorem 3.1.3, and it matches with Theorem 3.1.1.

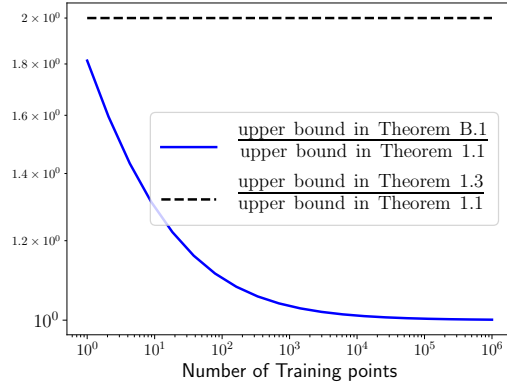


Figure B.1: Comparison between constants of Theorem 3.1.1, Theorem 3.1.3, and Theorem B.2.1 for the case $k \rightarrow \infty$.

B.3 Proofs of Section 3.2

Proof of Theorem 3.2.1. By the chain rule for the mutual information, we have

$$I(W; U^{(k)}, \tilde{Z}k) = I(W; \tilde{Z}^{(k)}) + I(W; U^{(k)} | \tilde{Z}k). \tag{B.2}$$

Since S is $\sigma(\tilde{Z}k, U^{(k)})$ -measurable, $I(W; U^{(k)}, \tilde{Z}k) = I(W; S, U^{(k)}, \tilde{Z}k)$. But then W is independent of $\tilde{Z}k, U^{(k)}$ given S , hence $I(W; S, U^{(k)}, \tilde{Z}k) = I(W; S)$. The result follows from the nonnegativity of mutual information. \square

Proof of Theorem 3.2.2. By Theorem 3.2.1, $I(W; U^{(k)} | \tilde{Z}k) = I(W; S) - I(W; \tilde{Z}k)$. Therefore, in order to prove the claim, we need to show $\lim_{k \rightarrow \infty} I(W; \tilde{Z}k) = 0$ when $I(W; S)$ is finite.

Recall that \mathcal{A} is a probability kernel from the space of tuples in \mathcal{Z} to \mathcal{W} . Assume $\mathcal{W} = \{w_1, \dots, w_m\}$. For each $l \in [m]$, let $\kappa_l(S) = \mathbb{P}^S[W = w_l]$ and $f_l : \mathcal{Z}^{kn} \rightarrow [0, 1]$ be a measurable function defined as

$$f_l(\tilde{Z}k) = \frac{1}{k^n} \sum_{u \in [k]^n} \kappa_l(\tilde{Z}k_u).$$

Letting $z, z' \in \mathcal{Z}^{kn}$ be two super-samples that only differ in one element, it is straightforward to see that

$$|f_l(z) - f_l(z')| \leq \frac{1}{k}.$$

Therefore, we can invoke McDiarmid's inequality to obtain

$$\mathbb{P}[|f_l(\tilde{Z}k) - \mathbb{E}[f_l(\tilde{Z}k)]| \geq \epsilon] \leq \exp\left(-\frac{2k\epsilon^2}{n}\right). \quad (\text{B.3})$$

Also, we have $\mathbb{E}[f_l(\tilde{Z}k)] = \mathbb{P}[W = w_l]$ as each element of $\tilde{Z}k$ is IID. Hence, $f_l(\tilde{Z}k) \rightarrow \mathbb{P}[W = w_l]$ in probability as k diverges.

By the definition of mutual information and KL divergence,

$$\begin{aligned} I(W; \tilde{Z}k) &= \mathbb{E}[\text{KL}(\mathbb{P}^{\tilde{Z}k}[W] \parallel \mathbb{P}[W])] \\ &= \mathbb{E}\left[\text{KL}\left(\frac{1}{k^n} \sum_{u \in [k]^n} \mathbb{P}^{\tilde{Z}k_u}[W] \parallel \mathbb{P}[W]\right)\right] \\ &= \mathbb{E}\left[\sum_{l=1}^m \frac{1}{k^n} \sum_{u \in [k]^n} \kappa_l(\tilde{Z}k_u) \log \frac{\frac{1}{k^n} \sum_{u \in [k]^n} \kappa_l(\tilde{Z}k_u)}{\mathbb{P}[W = w_l]}\right] \\ &= \sum_{l=1}^m \mathbb{E}\left[f_l(\tilde{Z}k) \log \frac{f_l(\tilde{Z}k)}{\mathbb{P}[W = w_l]}\right]. \end{aligned} \quad (\text{B.4})$$

Defining $\phi_l : [0, 1] \rightarrow \mathbb{R}$ as $\phi_l(x) = x \log \frac{x}{\mathbb{P}[W = w_l]}$, we have established

$$I(W; \tilde{Z}k) = \sum_{l=1}^m \mathbb{E}[\phi_l(f_l(\tilde{Z}k))]. \quad (\text{B.5})$$

Note that ϕ_l is a continuous and bounded function. By a standard result [Dur19, Thm. 2.3.4], $f_l(\tilde{Z}k) \rightarrow \mathbb{P}[W = w_l]$ in probability implies that

$$\mathbb{E}[\phi_l(f_l(\tilde{Z}k))] \rightarrow \mathbb{E}[\phi_l(\mathbb{P}[W = w_l])] = 0,$$

as k goes to infinity. Using this, we conclude that $I(W; \tilde{Z}k) \rightarrow 0$ as k diverges, as was to be shown. \square

Proof of Theorem B.2.1. For any $k \in \mathbb{N}$ define

$$\rho_i^{(k)}(m) = \begin{cases} -1 & \text{if } m = i, \\ \frac{1}{k-1} & \text{otherwise} \end{cases}.$$

where m and $i \in [k]$. Consider random variables $\tilde{Z}k$, U , and W as in the definition of $\text{CMI}_{\mathcal{D}}(\mathcal{A})$. Also, let $\hat{U} \stackrel{d}{=} U$ and $\hat{U} \perp\!\!\!\perp (\tilde{Z}k, W, U)$. Let $f : \mathcal{Z}^{kn} \times [k]^n \times \mathcal{W} \rightarrow [-1, 1]$

be given by

$$f(\tilde{z}^{(k)}, u, w) = \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^k \rho_i^{(k)}(u_j) \ell(w, z_{i,j}).$$

Then, by the Donsker–Varadhan variational formula [BLM13, Prop. 4.15] of the KL divergence we obtain

$$\begin{aligned} \text{CMI}_{\mathcal{D}}(\mathcal{A}) &= \mathbb{E}[\text{KL}(\mathbb{P}^{\tilde{Z}k, W}[U] \parallel \mathbb{P}[U])] \\ &\geq \lambda \mathbb{E}[f(\tilde{Z}k, U, W)] - \mathbb{E} \log \mathbb{E}^{\tilde{Z}k, W}[\exp(\lambda f(\tilde{Z}k, \hat{U}, W))] \end{aligned} \quad (\text{B.6})$$

It is straightforward to see that the first term in the RHS of Eq. (B.6) is $\mathbb{E}[f(\tilde{Z}k, U, W)] = \text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$. In what follows we analyze the second term in the RHS of Eq. (B.6). We begin with

$$\mathbb{E} \log \mathbb{E}^{\tilde{Z}k, W}[\exp(\lambda f(\tilde{Z}k, \hat{U}, W))] = \mathbb{E} \log \mathbb{E}^{\tilde{Z}k, W} \left[\prod_{j=1}^n \exp\left(\frac{\lambda}{n} \sum_{i=1}^k \rho_i^{(k)}(\hat{U}_j) \ell(W, Z_{i,j})\right) \right] \quad (\text{B.7})$$

$$\begin{aligned} &= \mathbb{E} \log \prod_{j=1}^n \mathbb{E}^{\tilde{Z}k, W} \exp\left(\frac{\lambda}{n} \sum_{i=1}^k \rho_i^{(k)}(\hat{U}_j) \ell(W, Z_{i,j})\right) \\ &\leq \mathbb{E} \sum_{j=1}^n \left(\exp\left(\frac{\lambda}{n}\right) - \frac{\lambda}{n} - 1 \right) \mathbb{E}^{\tilde{Z}k, W} \left[\sum_{i=1}^k \rho_i^{(k)}(\hat{U}_j) \ell(W, Z_{i,j}) \right]^2 \end{aligned} \quad (\text{B.8})$$

The first step follows from the independence of the elements in \hat{U} . The last inequality is obtained by the Bennet’s inequality [BLM13, Thm. 2.9] on the moment generating function and the fact that the elements of \hat{U} are independent of $(\tilde{Z}k, W)$. Also, we have used $\mathbb{E}^{\tilde{Z}k, W} \sum_{i=1}^k \rho_i^{(k)}(\hat{U}_j) \ell(W, Z_{i,j}) = 0$ since $\mathbb{E} \rho_i^{(k)}(\hat{U}_j) = 0$

and $|\sum_{i=1}^k \rho_i^{(k)}(\hat{U}_j)\ell(W, Z_{i,j})| \leq 1$ a.s.. For a fixed j , from Eq. (B.8) we obtain

$$\mathbb{E}\left[\sum_{i=1}^k \rho_i^{(k)}(\hat{U}_j)\ell(W, Z_{i,j})\right]^2 = \frac{1}{k}\mathbb{E}\sum_{\tilde{i}=[k]} \left[\frac{1}{k-1}\sum_{i \in [k], i \neq \tilde{i}} \ell(W, Z_{i,j}) - \ell(W, Z_{\tilde{i},j})\right]^2 \quad (\text{B.9})$$

$$= \frac{1}{k^2}\mathbb{E}\left[\sum_{u_j \in [k]} \mathbb{E}^{\tilde{Z}k, U_j=u_j} \sum_{\tilde{i}=[k]} \left[\frac{1}{k-1}\sum_{i \in [k], i \neq \tilde{i}} \ell(W, Z_{i,j}) - \ell(W, Z_{\tilde{i},j})\right]^2\right] \quad (\text{B.10})$$

$$= \frac{1}{k^2}\mathbb{E}\left[\sum_{u_j \in [k]} \mathbb{E}^{\tilde{Z}k, U_j=u_j} \left[\sum_{\tilde{i}=[k], \tilde{i} \neq u_j} \left[\frac{1}{k-1}\ell(W, Z_{u_j,j}) + \frac{1}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j}) - \ell(W, Z_{\tilde{i},j})\right]^2 + \left[\frac{1}{k-1}\sum_{i \in [k], i \neq \tilde{i}} \ell(W, Z_{i,j}) - \ell(W, Z_{u_j,j})\right]^2\right]\right] \quad (\text{B.11})$$

$$\leq \frac{1}{k^2}\mathbb{E}\left[\sum_{u_j \in [k]} \mathbb{E}^{\tilde{Z}k, U_j=u_j} \left[\sum_{\tilde{i}=[k], \tilde{i} \neq u_j} \left[\left(\frac{1}{k-1} + \frac{1}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j})\right)^2 + \ell(W, Z_{\tilde{i},j})^2 - 2\ell(W, Z_{\tilde{i},j})\frac{1}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j})\right] + \left[\frac{1}{k-1}\sum_{i \in [k], i \neq \tilde{i}} \ell(W, Z_{i,j})\right]^2 + 1\right]\right] \quad (\text{B.12})$$

$$= \frac{1}{k^2}\sum_{u_j \in [k]} \mathbb{E}\left[\sum_{\tilde{i} \in [k], u_j \neq \tilde{i}} \mathbb{E}^W \left[\left(\frac{1}{k-1} + \frac{1}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j})\right)^2 + \ell(W, Z_{\tilde{i},j})^2 - 2\ell(W, Z_{\tilde{i},j})\frac{1}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j})\right] + \mathbb{E}^W \left[\frac{1}{k-1}\sum_{i \in [k], i \neq \tilde{i}} \ell(W, Z_{i,j})\right]^2 + 1\right]\right] \quad (\text{B.13})$$

Here, Eq. (B.9) is obtained by taking the expectation over \hat{U}_j , the definition of $\rho_i^{(k)}(\hat{U}_j)$, and $\hat{U} \perp\!\!\!\perp (\tilde{Z}k, W)$. Then, Eq. (B.10) is by the law of iterated expectations. Specifically, we condition on U_j , and recall that based on the Definition 3.1.2 the j -th training sample is $Z_{U_j,j}$. Step Eq. (B.11) is by some manipulations. Eq. (B.12) is obtained by

$$\left[\frac{1}{k-1}\ell(W, Z_{u_j,j}) + \frac{1}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j}) - \ell(W, Z_{\tilde{i},j})\right]^2 \leq \left(\frac{1}{k-1} + \frac{1}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j})\right)^2 + \ell(W, Z_{\tilde{i},j})^2 - \ell(W, Z_{\tilde{i},j})\frac{2}{k-1}\sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j}),$$

and

$$\left[\frac{1}{k-1} \sum_{i \in [k], i \neq \tilde{i}} \ell(W, Z_{i,j}) - \ell(W, Z_{u_j,j}) \right]^2 \leq 1 + \left(\frac{1}{k-1} \sum_{i \in [k], i \neq \tilde{i}=u_j} \ell(W, Z_{i,j}) \right)^2.$$

Finally last step is obtained by changing the order of the expectation over W and $\tilde{Z}k$.

Then, we can simplify Eq. (B.13) by considering

$$\begin{aligned} & \mathbb{E}^W \left[\left(\frac{1}{k-1} + \frac{1}{k-1} \sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j}) \right)^2 + \ell(W, Z_{\tilde{i},j})^2 - 2\ell(W, Z_{\tilde{i},j}) \frac{1}{k-1} \sum_{i \in [k], i \neq \{\tilde{i}, u_j\}} \ell(W, Z_{i,j}) \right] \\ & \leq R_{\mathcal{D}}(W)^2 \frac{-k^2 + k + 2}{(k-1)^2} + R_{\mathcal{D}}(W) \frac{k^2 + k - 5}{(k-1)^2} + \frac{1}{(k-1)^2} \triangleq A_1(k, R_{\mathcal{D}}(W)) \end{aligned} \quad (\text{B.14})$$

$$\begin{aligned} & \mathbb{E}^W \left[\left[\frac{1}{k-1} \sum_{i \in [k], i \neq \tilde{i}} \ell(W, Z_{i,j}) \right]^2 + 1 \right] \leq R_{\mathcal{D}}(W)^2 \frac{k-2}{k-1} + R_{\mathcal{D}}(W) \frac{1}{k-1} + 1 \triangleq A_2(k, R_{\mathcal{D}}(W)). \end{aligned} \quad (\text{B.15})$$

Note that in Eq. (B.14) and Eq. (B.15), W and $Z_{i,j}$ s are independent, therefore $\mathbb{E}^W[\ell(W, Z_{i,j})] = R_{\mathcal{D}}(W)$. Also, we have $\text{Var}^W[\ell(W, Z_{i,j})] \leq R_{\mathcal{D}}(W)(1 - R_{\mathcal{D}}(W))$ because Bernoulli random variable has the largest variance among the $[0, 1]$ -bounded random variables with the same mean. Plugging Eq. (B.14) and Eq. (B.15) into Eq. (B.13) we obtain

$$\mathbb{E} \left[\sum_{i=1}^k \rho_i^{(k)}(U_j) \ell(W, Z_{i,j}) \right]^2 \leq \frac{1}{k} \mathbb{E}[(k-1)A_1(k, R_{\mathcal{D}}(W)) + A_2(k, R_{\mathcal{D}}(W))]. \quad (\text{B.16})$$

We can upper bound the LHS of Eq. (B.16) by maximizing it over $R_{\mathcal{D}}(W)$ to obtain

$$\frac{\partial[(k-1)A_1(k, R) + A_2(k, R)]}{\partial R} = 0 \Rightarrow R^* = \frac{k^2 + k - 4}{2k^2 - 4k}.$$

We can plug the expression for R^* into Eq. (B.16) to get

$$\begin{aligned} & \mathbb{E} \left[\sum_{i=1}^k \rho_i^{(k)}(U_j) \ell(W, Z_{i,j}) \right]^2 \leq \frac{1}{k} \mathbb{E}[(k-1)A_1(k, R^*) + A_2(k, R^*)] \\ & = \frac{k^3 + 7k^2 - 8k - 16}{4(k^3 - 2k^2)}. \end{aligned} \quad (\text{B.17})$$

Then, plugging Eq. (B.17) into Eq. (B.6) yields

$$\inf_{\lambda \geq 0} \frac{\text{CMI}_{\mathcal{D}}(\mathcal{A})}{\lambda} + \frac{\exp(\frac{\lambda}{n}) - \frac{\lambda}{n} - 1}{\frac{\lambda}{n}} \frac{k^3 + 7k^2 - 8k - 16}{4(k^3 - 2k^2)} \geq \text{EGE}_{\mathcal{D}}(\mathcal{A}_n). \quad (\text{B.18})$$

Finally, the closed form solution of Eq. (B.18) is given by

$$\begin{aligned} & \frac{\partial \left[\frac{\text{CMI}_{\mathcal{D}}(\mathcal{A})}{\lambda} + \frac{\exp(\frac{\lambda}{n}) - \frac{\lambda}{n} - 1}{\frac{\lambda}{n}} \frac{k^3 + 7k^2 - 8k - 16}{4(k^3 - 2k^2)} \right]}{\partial \lambda} = 0 \Rightarrow \\ & \lambda^* = n \mathfrak{W} \left(\left(\frac{4(k^3 - 2k^2) \text{CMI}_{\mathcal{D}}(\mathcal{A})}{n(k^3 + 7k^2 - 8k - 16)} - 1 \right) \exp(-1) \right) + n, \end{aligned}$$

which is the desired result. \square

B.4 Proofs of Section 6.4

Proof of Theorem 3.3.1. With $k = 2$, recall from the introduction

$$\tilde{Z}2 = \begin{pmatrix} Z_{1,1} & \cdots & Z_{1,n} \\ Z_{2,1} & \cdots & Z_{2,n} \end{pmatrix} \sim \mathcal{D}^{\otimes 2n},$$

and $U = (U_1, \dots, U_n) \in \{1, 2\}^n$ where U_i s are IID, and the marginal distribution follows $U_i \sim \text{Bern}(\frac{1}{2})$ for $i \in [n]$. Furthermore, recall $S = \{Z_{U_1,1}, \dots, Z_{U_n,n}\}$. The expected generalization error can be written as

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (-1)^{U_i} (\ell(Z_{1,i}, W) - \ell(Z_{2,i}, W)) \right] \quad (\text{B.19})$$

$$= \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m (-1)^{U_{J_i}} (\ell(Z_{1,J_i}, W) - \ell(Z_{2,J_i}, W)) \right], \quad (\text{B.20})$$

where the last equality follows because J is independent of U_J , $\tilde{Z}2$, and W .

Define \tilde{W} , \tilde{U}_J , and \tilde{J} such that $(W, U_J, J, \tilde{Z}2) \stackrel{d}{=} (\tilde{W}, \tilde{U}_J, \tilde{J}, \tilde{Z}2)$, and \tilde{W} , \tilde{U}_J , and \tilde{J} are independent given $\tilde{Z}2$. By the Donsker–Varadhan variational formula [BLM13, Prop. 4.15] and the disintegration theorem [Kal06, Thm. 6.4], for all measurable functions g in \mathcal{G} , i.e., the class of all functions g such that

$(\mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{W}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{U}_J] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{J}]) (\exp g) < \infty$, with probability one we have

$$I^{\tilde{Z}^{(2)}}(W, J; U_J) = \text{KL}(\mathbb{P}^{\tilde{Z}^{(2)}}[W, J, U_J] \parallel \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{W}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{J}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{U}_J]) \quad (\text{B.21})$$

$$= \sup_{g \in \mathcal{G}} \mathbb{P}^{\tilde{Z}^{(2)}}[W, J, U_J](g) - \log \left[\left(\mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{W}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{J}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{U}_J] \right) (\exp g) \right]. \quad (\text{B.22})$$

Define $f(w, j, u_j) \triangleq \frac{\lambda}{m} \sum_{i=1}^m (-1)^{u_{j_i}} (\ell(z_{1,j_i}, w) - \ell(z_{2,j_i}, w))$ where $\lambda \geq 0$. By Eq. (B.22), we can write

$$I^{\tilde{Z}^{(2)}}(W, J; U_J) \geq \mathbb{P}^{\tilde{Z}^{(2)}}[W, U_J, J](f) - \log \left[\left(\mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{W}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{U}_J] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{J}] \right) (\exp f) \right]. \quad (\text{B.23})$$

Considering the second term of the RHS of Eq. (B.23), Hoeffding's lemma implies that

$$\left(\mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{W}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{U}_J] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{J}] \right) (\exp f) \quad (\text{B.24})$$

$$= \mathbb{E}^{\tilde{Z}^{(2)}} \exp \left(\frac{\lambda}{m} \sum_{i=1}^m (-1)^{\tilde{U}_{J_i}} \left(\ell \left(Z_{1,\tilde{J}_i}, \tilde{W} \right) - \ell \left(Z_{2,\tilde{J}_i}, \tilde{W} \right) \right) \right) \quad (\text{B.25})$$

$$= \mathbb{E}^{\tilde{Z}^{(2)}} \mathbb{E}^{\tilde{Z}^{(2)}, \tilde{W}, \tilde{J}} \prod_{i=1}^m \exp \left(\frac{\lambda}{m} (-1)^{\tilde{U}_{J_i}} \left(\ell \left(Z_{1,\tilde{J}_i}, \tilde{W} \right) - \ell \left(Z_{2,\tilde{J}_i}, \tilde{W} \right) \right) \right) \quad (\text{B.26})$$

$$\leq \mathbb{E}^{\tilde{Z}^{(2)}} \mathbb{E}^{\tilde{Z}^{(2)}, \tilde{W}, \tilde{J}} \prod_{i=1}^m \exp \left(\frac{\lambda^2 \left(\ell \left(Z_{1,\tilde{J}_i}, \tilde{W} \right) - \ell \left(Z_{2,\tilde{J}_i}, \tilde{W} \right) \right)^2}{2m^2} \right) \quad (\text{B.27})$$

$$\leq \exp \left(\frac{\lambda^2}{2m} \right), \quad (\text{B.28})$$

where we use the fact that

$$\left(\mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{W}] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{U}_J] \otimes \mathbb{P}^{\tilde{Z}^{(2)}}[\tilde{J}] \right) (f) = 0.$$

Substituting the bound in Eq. (B.28) into Eq. (B.23), rearranging and taking expectations, we obtain

$$\mathbb{E} \frac{1}{m} \sum_{i=1}^m (-1)^{U_{J_i}} (\ell(Z_{1,J_i}, W) - \ell(Z_{2,J_i}, W)) \leq \mathbb{E} \inf_{\lambda \geq 0} \frac{I^{\tilde{Z}^{(2)}}(W, J; U_J) + \frac{\lambda^2}{2m}}{\lambda} \quad (\text{B.29})$$

$$= \mathbb{E} \sqrt{\frac{2}{m} I^{\tilde{Z}^{(2)}}(W, J; U_J)}. \quad (\text{B.30})$$

Moreover, we have a.s.

$$I^{\tilde{Z}2}(J, W; U_J) - \underbrace{I^{\tilde{Z}2}(J; U_J)}_0 = I^{\tilde{Z}2}(W; U_J|J). \quad (\text{B.31})$$

Here, $I^{\tilde{Z}2}(J; U_J) = 0$ since J is independent of U_J given $\tilde{Z}2$. Plugging Eq. (B.31) into Eq. (B.30), we obtain the desired result. \square

Proof of Theorem 3.3.2. Consider

$$I(W; U_{J^{(m_1)}}|\tilde{Z}2, J^{(m_1)}) = \text{H}(U_{J^{(m_1)}}|J^{(m_1)}, \tilde{Z}2) - \text{H}(U_{J^{(m_1)}}|J^{(m_1)}, \tilde{Z}2, W) \quad (\text{B.32})$$

$$= \frac{1}{\binom{n}{m_1}} \sum_{K_1 \in [n]_{m_1}} \text{H}(U_{K_1}|\tilde{Z}2) - \text{H}(U_{J^{(m_1)}}|J^{(m_1)}, \tilde{Z}2, W) \quad (\text{B.33})$$

$$= \frac{1}{\binom{n}{m_1}} \sum_{K_1 \in [n]_{m_1}} \text{H}(U_{K_1}|\tilde{Z}2) - \frac{1}{\binom{n}{m_1}} \sum_{K_1 \in [n]_{m_1}} \text{H}(U_{K_1}|\tilde{Z}2, W). \quad (\text{B.34})$$

Eq. (B.33) follows because $\tilde{Z}2 \perp\!\!\!\perp J^{(m_1)}$, while Eq. (B.34) follows because the event $\{J^{(m_1)} = K_1\}$ is independent of $(W, U_{K_1}, \tilde{Z}2)$. Then

$$\frac{1}{m_1} I(W; U_{J^{(m_1)}}|\tilde{Z}2, J^{(m_1)}) = \frac{1}{m_1 \binom{n}{m_1}} \sum_{K_1 \in [n]_{m_1}} [\text{H}(U_{K_1}) - \text{H}(U_{K_1}|W, \tilde{Z}2)] \quad (\text{B.35})$$

$$= \frac{1}{n} \text{H}(U) - \frac{1}{m_1 \binom{n}{m_1}} \sum_{K_1 \in [n]_{m_1}} \text{H}(U_{K_1}|W, \tilde{Z}2) \quad (\text{B.36})$$

$$= \frac{1}{m_2 \binom{n}{m_2}} \sum_{K_2 \in [n]_{m_2}} \text{H}(U_{K_2}) - \frac{1}{m_1 \binom{n}{m_1}} \sum_{K_1 \in [n]_{m_1}} \text{H}(U_{K_1}|W, \tilde{Z}2) \quad (\text{B.37})$$

$$\leq \frac{1}{m_2 \binom{n}{m_2}} \sum_{K_2 \in [n]_{m_2}} [\text{H}(U_{K_2}) - \text{H}(U_{K_2}|W, \tilde{Z}2)] \quad (\text{B.38})$$

$$= \frac{1}{m_2} I(W; U_{J^{(m_2)}}|\tilde{Z}2, J^{(m_2)}). \quad (\text{B.39})$$

Eq. (B.35) follows from Eq. (B.34) and the fact that $U \perp\!\!\!\perp \tilde{Z}2$, while Eq. (B.36) follows from each element of U being IID. Eq. (B.38) follows from Lemma B.6.1, which is a modified version of the Han's inequality [Te 78]. Finally, the last step follows from using the same line of reasoning as in Eq. (B.32) to Eq. (B.34).

Having established Eq. (3.4), the claim follows from

$$\mathbb{E} \sqrt{\frac{2}{m} I^{\tilde{Z}2}(W; U_J | J)} \leq \sqrt{\frac{2}{m} I(W; U_J | \tilde{Z}2, J)} \quad (\text{B.40})$$

$$\leq \sqrt{\frac{2}{n} I(W; U | \tilde{Z}2)}, \quad (\text{B.41})$$

where Eq. (B.40) is Jensen's inequality, and Eq. (B.41) is the direct application of Eq. (3.4) with $m_1 = m$ and $m_2 = n$. This proves the desired result. \square

Proof of Theorem 3.3.4. By the Donsker–Varadhan variational formula [BLM13, Prop. 4.15] and the disintegration theorem [Kal06, Thm. 6.4], with probability one, for all measurable functions g such that $(\mathbb{P}^{\tilde{Z}2}[\tilde{W}] \otimes \mathbb{P}^{\tilde{Z}2}[\tilde{U}_i])(\exp g) < \infty$, we have

$$I^{\tilde{Z}2}(U_i, W) = \text{KL}(\mathbb{P}^{\tilde{Z}2}[U_i, W] \parallel \mathbb{P}^{\tilde{Z}2}[\tilde{U}_i] \otimes \mathbb{P}^{\tilde{Z}2}[\tilde{W}]) \quad (\text{B.42})$$

$$\geq \mathbb{P}^{\tilde{Z}2}[U_i, W][g(W, \tilde{Z}2, U_i)] - \log \mathbb{P}^{\tilde{Z}2}[\tilde{U}_i] \otimes \mathbb{P}^{\tilde{Z}2}[\tilde{W}][\exp(g(\tilde{W}, \tilde{Z}2, \tilde{U}_i))] \quad (\text{B.43})$$

where $(W, U_i, \tilde{Z}2) \stackrel{d}{=} (\tilde{W}, \tilde{U}_i, \tilde{Z}2)$ and $\tilde{W} \perp\!\!\!\perp \tilde{U}_i \mid \tilde{Z}2$. For $i \in [n]$, let

$$g_i(W, \tilde{Z}2, U_i) \triangleq \lambda (-1)^{U_i} (\ell(Z_{1,i}, W) - \ell(Z_{2,i}, W)).$$

Hoeffding's lemma implies that

$$\mathbb{P}^{\tilde{Z}2}[\tilde{U}_i] \otimes \mathbb{P}^{\tilde{Z}2}[\tilde{W}][\exp(g_i(\tilde{W}, \tilde{Z}2, \tilde{U}_i))] \leq \exp\left(\frac{\lambda^2}{2}\right), \quad (\text{B.44})$$

where in the last line we have used $g_i \in [-\lambda, \lambda]$ a.s. From (B.43), we obtain

$$\mathbb{E}^{\tilde{Z}2} (-1)^{U_i} (\ell(Z_{1,i}, W) - \ell(Z_{2,i}, W)) \leq \inf_{\lambda \geq 0} \frac{\text{KL}(\mathbb{P}^{\tilde{Z}2}[U_i, W] \parallel \mathbb{P}^{\tilde{Z}2}[U_i] \otimes \mathbb{P}^{\tilde{Z}2}[W]) + \frac{\lambda^2}{2}}{\lambda} \quad (\text{B.45})$$

$$= \sqrt{2I^{\tilde{Z}2}(W; U_i)}. \quad (\text{B.46})$$

Then, averaging over i and taking expectations,

$$\mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] = \mathbb{E} \frac{1}{n} \sum_{i=1}^n \mathbb{E}^{\tilde{Z}^2} (-1)^{U_i} (\ell(Z_{1,i}, W) - \ell(Z_{2,i}, W)) \quad (\text{B.47})$$

$$\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \sqrt{2I^{\tilde{Z}^2}(W; U_i)}. \quad (\text{B.48})$$

□

Proof of Theorem 3.3.7. For any two random measures $P(\tilde{Z}^2, U_{J^c}, J)$ and $Q(\tilde{Z}^2, U)$ on \mathcal{W} , the Donsker–Varadhan variational formula [BLM13, Prop. 4.15] and the disintegration theorem [Kal06, Thm. 6.4], give that with probability one

$$\text{KL}(Q(\tilde{Z}^2, U) \parallel P(\tilde{Z}^2, U_{J^c}, J)) = \sup_{g \in \mathcal{G}} \left(Q(\tilde{Z}^2, U) [g] - \log P(\tilde{Z}^2, U_{J^c}, J) [\exp g] \right) \quad (\text{B.49})$$

where $\mathcal{G} = \{g : P(\tilde{Z}^2, U_{J^c}, J)(\exp g) < \infty\}$.

Let $g = \frac{\lambda}{m} \sum_{j \in J} (-1)^{U_j} (\ell(Z_{1,j}, W) - \ell(Z_{2,j}, W))$. First, note that

$$\mathbb{E}^{\tilde{Z}^2, U_{J^c}, J} \left[\frac{\lambda}{m} \sum_{j \in J} (-1)^{U_j} (\ell(Z_{1,j}, W) - \ell(Z_{2,j}, W)) \right] = 0.$$

This is because $\{U_j\}_{j \in J}$ are independent of \tilde{Z}^2, U_{J^c} , and J . Moreover, g is $[-\lambda, \lambda]$ -bounded. Therefore, we can use the Hoeffding’s lemma to obtain

$$\log P(\tilde{Z}^2, U_{J^c}, J) (\exp g) \leq \frac{\lambda^2}{2}.$$

Hence, from Eq. (B.49), we conclude that

$$\begin{aligned} & Q(\tilde{Z}^2, U) \left[\frac{1}{m} \sum_{j \in J} (-1)^{U_j} (\ell(Z_{1,j}, W) - \ell(Z_{2,j}, W)) \right] \\ & \leq \inf_{\lambda > 0} \frac{\text{KL}(Q(\tilde{Z}^2, U, J) \parallel P(\tilde{Z}^2, U_{J^c}, J))}{\lambda} + \frac{\lambda}{2} = \sqrt{2\text{KL}(Q(\tilde{Z}^2, U) \parallel P(\tilde{Z}^2, U_{J^c}, J))} \end{aligned}$$

almost surely. Finally, since $J \perp\!\!\!\perp (\tilde{Z}^2, U)$ we get

$$\begin{aligned} & Q(\tilde{Z}^2, U) \left[\frac{1}{m} \sum_{j \in J} (-1)^{U_j} (\ell(Z_{1,j}, W) - \ell(Z_{2,j}, W)) \right] \\ &= Q(\tilde{Z}^2, U) \left[\frac{1}{n} \sum_{i=1}^n (-1)^{U_i} (\ell(Z_{1,i}, W) - \ell(Z_{2,i}, W)) \right] \\ &= \mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] \end{aligned}$$

The desired result follows. \square

B.5 Proofs of Section 3.4

Proof of Theorem 3.4.2. Considering the generalization bound in Theorem 3.3.7 and Lemma 3.3.6, we can write

$$\begin{aligned} \mathbb{E} \left[R_{\mathcal{D}}(W) - \hat{R}_S(W) \right] &\leq \mathbb{E} \sqrt{2\text{KL}(Q_T(S) \parallel P_T(\tilde{Z}^2, U_{J^c}, J))} \\ &\leq \mathbb{E} \sqrt{\sum_{t=1}^T 2\mathbb{E}^{\tilde{Z}^2, U, J} \text{KL}(Q_t \parallel P_t)}. \end{aligned} \quad (\text{B.50})$$

First, note that from Eq. (3.11) it follows that

$$Q_t = \mathcal{N}(\mu_{Q_t}, \frac{2\eta_t}{\beta_t} \mathbb{I}_d),$$

where the mean is given by

$$\mu_{Q_t} = W_{t-1} - \eta_{t-1} \frac{n-1}{n} \nabla \tilde{R}_{S_{J^c}}(W_{t-1}) - \frac{\eta_{t-1}}{n} \left(\mathbb{1}[U_J = 1] \nabla \tilde{\ell}(Z_{1,J}, W_{t-1}) + \mathbb{1}[U_J = 2] \nabla \tilde{\ell}(Z_{2,J}, W_{t-1}) \right).$$

Next, we propose the following construction of P_t . Note that P_t is \mathcal{F}_t -measurable random probability measure where

$$\mathcal{F}_t = \sigma(S_{J^c}, Z_{1,J}, Z_{2,J}, J, W_{0:t-1}).$$

Hence we can exploit the information in the trajectory up to time t to construct P_t . In particular, we use the information in \mathcal{F}_t to perform a binary hypothesis testing in

which the two hypotheses are defined as

$$\begin{aligned}\mathcal{H}_1 &: U_J = 1, \\ \mathcal{H}_2 &: U_J = 2.\end{aligned}$$

Equivalently, \mathcal{H}_1 and \mathcal{H}_2 can also be described as the hypotheses that $Z_{1,J}$ is a member of the training set and $Z_{2,J}$ is a member of the training set, respectively. Denote $\pi_t = (\pi_{t,1}, \pi_{t,2})$ as a probability vector whose i -th element shows the belief of the prior at time t that the true hypothesis is \mathcal{H}_i for $i \in \{1, 2\}$. Then, we consider the prior as

$$P_{t|} = \mathcal{N}(\mu_{P_{t|}}, \frac{2\eta_{t-1}}{\beta_{t-1}} \mathbb{I}_d), \quad (\text{B.51})$$

where

$$\mu_{P_{t|}} = W_{t-1} - \eta_{t-1} \frac{n-1}{n} \nabla \tilde{R}_{S_{J^c}}(W_{t-1}) - \frac{\eta_{t-1}}{n} \left(\pi_{t,1} \nabla \tilde{\ell}(Z_{1,J}, W_{t-1}) + \pi_{t,2} \nabla \tilde{\ell}(Z_{2,J}, W_{t-1}) \right). \quad (\text{B.52})$$

Here $\pi_1 = (\frac{1}{2}, \frac{1}{2})$. Then, we construct the belief vector π_t for $t \geq 2$ using the log-likelihood ratio as

$$\pi_t = \left(\theta \left(\log \frac{\mathbb{P}^{\mathcal{F}_t}[\mathcal{H}_1]}{\mathbb{P}^{\mathcal{F}_t}[\mathcal{H}_2]} \right), 1 - \theta \left(\log \frac{\mathbb{P}^{\mathcal{F}_t}[\mathcal{H}_1]}{\mathbb{P}^{\mathcal{F}_t}[\mathcal{H}_2]} \right) \right), \quad (\text{B.53})$$

where $\theta : \mathbb{R} \rightarrow [0, 1]$. Also, we might expect that the optimal θ satisfies $\theta(0) = \frac{1}{2}$, $\lim_{x \rightarrow \infty} \theta(x) = 1$, and $\lim_{x \rightarrow -\infty} \theta(x) = 0$.

Denote probability density function $\mathbb{P}^{\tilde{Z}_2, U_{J^c}, \mathcal{H}_k, W_0}[W_{1:t-1}]$ as $f_k(W_{1:t-1})$ for $k \in \{1, 2\}$. Due to Markov structure of the update rule in Eq. (3.11), we have

$$\begin{aligned}f_k(W_{1:t-1}) &= \\ \prod_{i=1}^{t-1} \left(\frac{\beta_{i-1}}{4\pi\eta_{i-1}} \right)^{\frac{d}{2}} \exp \left(- \frac{\beta_{i-1} \|W_i - W_{i-1} + \eta_{i-1} \frac{n-1}{n} \nabla \tilde{R}_{S_{J^c}}(W_{i-1}) + \frac{\eta_{i-1}}{n} \nabla \tilde{\ell}(Z_{k,J}, W_{i-1})\|^2}{4\eta_{i-1}} \right).\end{aligned} \quad (\text{B.54})$$

Here, Eq. (B.54) is obtained by the Markov property of the update rule in Eq. (3.11). Then, since the prior distribution on \mathcal{H}_1 and \mathcal{H}_2 is uniform, we have

$$\log \frac{\mathbb{P}^{\mathcal{F}_t}[\mathcal{H}_1]}{\mathbb{P}^{\mathcal{F}_t}[\mathcal{H}_2]} = \log \frac{f_1(W_{1:t-1})}{f_2(W_{1:t-1})} \quad (\text{B.55})$$

$$= Y_{t,2} - Y_{t,1}, \quad (\text{B.56})$$

where $Y_{t,1}$ and $Y_{t,2}$ are given by

$$\begin{aligned} Y_{t,1} &\triangleq \sum_{i=1}^{t-1} \frac{\beta_{i-1}}{4\eta_{i-1}} \|W_i - W_{i-1} + \eta_{i-1} \frac{n-1}{n} \nabla \tilde{R}_{S_{J^c}}(W_{i-1}) + \frac{\eta_{i-1}}{n} \nabla \tilde{\ell}(Z_{1,J}, W_{i-1})\|^2, \\ Y_{t,2} &\triangleq \sum_{i=1}^{t-1} \frac{\beta_{i-1}}{4\eta_{i-1}} \|W_i - W_{i-1} + \eta_{i-1} \frac{n-1}{n} \nabla \tilde{R}_{S_{J^c}}(W_{i-1}) + \frac{\eta_{i-1}}{n} \nabla \tilde{\ell}(Z_{2,J}, W_{i-1})\|^2. \end{aligned} \quad (\text{B.57})$$

Therefore, the belief vector is given by

$$\pi_t = \left(\theta((Y_{t,2} - Y_{t,1})), 1 - \theta((Y_{t,2} - Y_{t,1})) \right), \quad (\text{B.58})$$

where $Y_{0,1} = Y_{0,2} = 0$ and for $t \geq 2$, $Y_{t,1}$ and $Y_{t,2}$ are given by Eq. (B.57). To conclude the proof, we obtain

$$\text{KL}(Q_T(S) \parallel P_T(\tilde{Z}_2, U_{J^c}, J)) \leq \sum_{t=1}^T \mathbb{E}^{\tilde{Z}_2, U, J} \text{KL}(Q_t \parallel P_t) \quad (\text{B.59})$$

$$= \sum_{t=1}^T \mathbb{E}^{\tilde{Z}_2, U, J} \frac{\beta_{t-1} \eta_{t-1} \|(\mathbb{1}[U_J = 1] - \pi_{t,1}) \nabla \tilde{\ell}(Z_{1,J}, W_{t-1}) + (\mathbb{1}[U_J = 2] - \pi_{t,2}) \nabla \tilde{\ell}(Z_{2,J}, W_{t-1})\|^2}{4n^2} \quad (\text{B.60})$$

$$= \sum_{t=1}^T \mathbb{E}^{\tilde{Z}_2, U, J} \frac{\beta_{t-1} \eta_{t-1} (\mathbb{1}[U_J = 1] - \pi_{t,1})^2 \|\nabla \tilde{\ell}(Z_{1,J}, W_{t-1}) - \nabla \tilde{\ell}(Z_{2,J}, W_{t-1})\|^2}{4n^2} \quad (\text{B.61})$$

Finally, plugging Eq. (B.61) into Eq. (B.50), we get the desired result in Eq. (3.14). \square

B.6 Conditional Han's Inequality

Lemma B.6.1. *Let (X_1, \dots, X_n, Y) be $n + 1$ -dimensional random variable where X_1, \dots, X_n are discrete random variables. Then,*

$$\frac{1}{k \binom{n}{k}} \sum_{T \in [n]_k} \mathbb{H}(X_T|Y)$$

is decreasing in k .

Proof. For notational convenience, let $\bar{H}_k(X_{[n]}|Y) = \frac{1}{\binom{n}{k}} \sum_{T \in [n]_k} \mathbb{H}(X_T|Y)$. Note that if we manage to show that

$$\bar{H}_k(X_{[n]}|Y) - \bar{H}_{k-1}(X_{[n]}|Y) \leq \bar{H}_{k+1}(X_{[n]}|Y) - \bar{H}_k(X_{[n]}|Y), \quad (\text{B.62})$$

then the result in Lemma B.6.1 follows. To show Eq. (B.62), we can write

$$\begin{aligned} & \mathbb{H}(X_1, \dots, X_{k+1}|Y) + \mathbb{H}(X_1, \dots, X_{k-1}|Y) \\ &= \mathbb{H}(X_1, \dots, X_k|Y) + \mathbb{H}(X_{k+1}|X_1, \dots, X_k, Y) + \mathbb{H}(X_1, \dots, X_{k-1}|Y) \end{aligned} \quad (\text{B.63})$$

$$\leq \mathbb{H}(X_1, \dots, X_k|Y) + \mathbb{H}(X_{k+1}|X_1, \dots, X_{k-1}, Y) + \mathbb{H}(X_1, \dots, X_{k-1}|Y) \quad (\text{B.64})$$

$$= \mathbb{H}(X_1, \dots, X_k|Y) + \mathbb{H}(X_1, \dots, X_{k-1}, X_{k+1}|Y). \quad (\text{B.65})$$

Here in Eq. (B.64), we drop X_k from the condition in the second term. Therefore, we have

$$\mathbb{H}(X_1, \dots, X_{k+1}|Y) + \mathbb{H}(X_1, \dots, X_{k-1}|Y) \leq \mathbb{H}(X_1, \dots, X_k|Y) + \mathbb{H}(X_1, \dots, X_{k-1}, X_{k+1}|Y). \quad (\text{B.66})$$

Then, by averaging Eq. (B.66) over all $n!$ permutation of $\{1, \dots, n\}$, we get the desired result in Eq. (B.62). \square

B.7 Details of Experiments

In this section, we discuss the details behind the experiments as well as the details of minimizing the generalization bound in Theorem 3.4.2.

B.7.1 Network architectures and learning curve

Tables B.1 to B.4 summarize the hyper-parameters we used for the experiments. Also, in Fig. B.2 we plot the learning curves for the experiments reported in Section 3.4.2.

Dataset	MNIST
Architecture	MLP(784-500-500-10)
η_t	$0.06 \times (0.95)^{\lceil \frac{t}{50} \rceil}$
$\frac{2\eta_t}{\beta_t}$	$10^{-8} + (3 \times 10^{-6} - 10^{-8}) \times \exp(-0.5 \lceil \frac{t}{50} \rceil)$
Number of iterations	900
Final training error	$4.33 \pm 0.01\%$
Generalization error	$0.88 \pm 0.01\%$
Number of training examples	20000
Number of runs	100

Table B.1: Details of Experiments reported for MNIST with MLP

Dataset	MNIST
Architecture	CL($5 \times 5(32)$)-MaxPool(2×2)-CL($5 \times 5(64)$) MaxPool(2×2)-FC(128)-FC(10)
η_t	$0.05 \times (0.90)^{\lceil \frac{t}{40} \rceil}$
$\frac{2\eta_t}{\beta_t}$	$10^{-8} + (10^{-5} - 10^{-8}) \times \exp(-0.5 \lceil \frac{t}{40} \rceil)$
Number of iterations	700
Final training error	$2.59 \pm 0.01\%$
Generalization error	$0.55 \pm 0.01\%$
Number of training examples	20000
Number of runs	100

Table B.2: Details of Experiments reported for MNIST with CNN

Dataset	Fashion-FMNIST
Architecture	CL($5 \times 5(32)$)-MaxPool(2×2)-CL($5 \times 5(64)$) MaxPool(2×2)-FC(200)-FC(10)
η_t	$0.07 \times (0.95)^{\lceil \frac{t}{50} \rceil}$
$\frac{2\eta_t}{\beta_t}$	$5 \times 10^{-8} + (7 \times 10^{-6} - 5 \times 10^{-8}) \times \exp(-0.3 \lceil \frac{t}{50} \rceil)$
Number of iterations	1300
Final training error	$7.96 \pm 0.03\%$
Generalization error	$3.71 \pm 0.03\%$
Number of training examples	20000
Number of runs	100

Table B.3: Details of Experiments reported for Fashion-MNIST with CNN

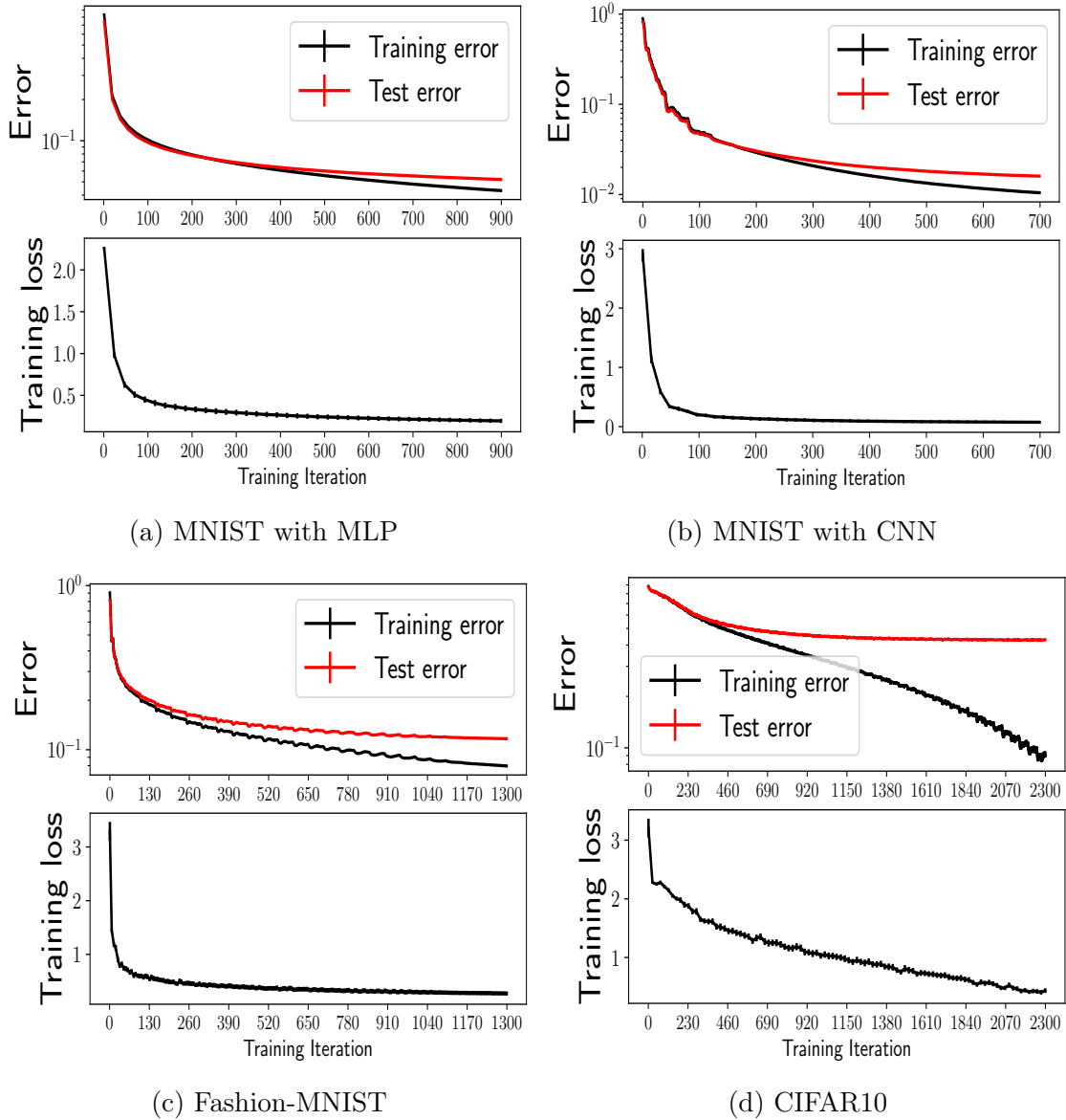


Figure B.2: Learning curves. These plots show the training error, error on the test set, and the training loss. The loss functions is cross-entropy. Note y-axes for the error plots are log-scale.

Dataset	CIFAR10
Architecture	CL(3 × 3(32))-MaxPool(2 × 2)-CL(3 × 3(64)) CL(3 × 3(32))-MaxPool(2 × 2)-FC(128)-FC(10)
η_t	$0.15 \times (0.98)^{\lceil \frac{t}{50} \rceil}$
$\frac{2\eta_t}{\beta_t}$	$10^{-9} + (3 \times 10^{-5} - 10^{-9}) \times \exp(-0.3 \lceil \frac{t}{50} \rceil)$
Number of iterations	2300
Final training error	$9.39 \pm 0.46\%$
Generalization error	$32.89 \pm 0.44\%$
Number of training examples	15000
Number of runs	100

Table B.4: Details of Experiments reported for CIFAR10 with CNN

B.7.2 Optimizing the bound over the choice of θ function

Our generalization bound in Theorem 3.4.2 consists of an infimum over the functions in Θ . To study the impact of infimum, we consider the family of functions Θ given by

$$\Theta = \{\theta_a(x) | \exists a > 0 \text{ such that } \theta_a(x) = \frac{1}{2}(1 + \operatorname{erf}(\frac{x}{a})) \text{ or } \theta_a(x) = \frac{1}{2}(1 + \tanh(\frac{x}{a}))\}.$$

Then, we divide the samples of the optimization trajectory into two sets of equal size: training set and the test set. Then, we optimize over a to find the $\theta_{a^*}(x)$ that achieves the minimum expected generalization over the training set. The numbers reported in Table 3.1 are based on the evaluation of $\theta_{a^*}(x)$ over the test set. Thus, the number reported in Table 3.1 are unbiased estimate of the generalization bound in Theorem 3.4.2.

C

Appendix of Chapter 4

C.1 Known Bounds for Learning VC Classes

In this part, we provide a landscape of the known results for the learning VC classes. One key distinction is proper learning versus improper learning. In particular, for every VC class with dimension d , there exists a consistent and improper learning algorithm that achieves $O(d/n)$ risk under realizability, and this bound is optimal [hanneke2016optimal; HLW94]. The situation for proper learning is much more complicated. In general, the achievable rate for the proper learning of VC classes is off by a log factor, i.e., $O(d \log(n)/n)$. Bousquet, Hanneke, Moran, and Zhivotovskiy [BHMZ20] show that when the dual Helly and hollow star number, which are combinatorial complexity measures of the class, agree, then they characterize the existence of an optimal proper learner. Also, a subclass of VC for which the log factor provably cannot be removed using proper learners is characterized in [BHMZ20, Thm. 11]. Moreover, for general ERM, Hanneke [Han16] shows the finiteness of star number is a necessary and sufficient condition under which we can remove the log factor using any arbitrary ERMs.

It is interesting to note that the results of moran2016sample reveal a connection between the general sample compression schemes and VC classes. However, it is not known whether, in general, the optimal rates for VC classes are always witnessed by compression schemes.

C.2 Proof of Theorem 4.2.3

Fix $n \in \mathbb{N}$. We have $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) = \mathbb{E}[I^Z(L; U)]$ by definition, and $I^Z(L; U) \leq H^Z(L)$ a.s., by the nonnegativity of (conditional) entropy. It then suffices to bound the cardinality, C , of the support of the conditional distribution of L , given Z , because $H^Z(L) \leq \log C$ a.s. For all $i \in \{0, 1\}$ and $j \in [n]$, write $Z_{i,j} = (X_{i,j}, Y_{i,j})$, and consider the set of possible predictions on Z ,

$$P = \{p \in \{0, 1\}^{\{0,1\} \times [n]} : \exists h \in \mathcal{H}_n, \forall i \in \{0, 1\}, j \in [n], p_{i,j} = h(X_{i,j})\}. \quad (\text{C.1})$$

The set P is precisely the set of possible labellings of the $2n$ inputs $(X_{i,j})$, which is bounded by the growth function of \mathcal{H}_n evaluated at $2n$ points. By the Sauer–Shelah lemma, the cardinality of P is thus bounded above by $6n^{dn}$. Note that the support of the conditional distribution of L given Z is

$$\{c \in \{0, 1\}^{\{0,1\} \times [n]} : \exists p \in P, \forall i \in \{0, 1\}, \forall j \in [n], \text{ if and only if } c_{i,j} = \mathbb{1}[p_{i,j} \neq Y_{i,j}]\}. \quad (\text{C.2})$$

Therefore, the cardinality of the support is no greater than that of P , hence $C \leq 6n^{dn}$.

C.3 Proof of Theorem 4.4.4

We prove the claim by contradiction. Pick f and $c \geq 0$. Let \mathcal{H} be a concept class with finite VC dimension d as shown to exist by Theorem 4.4.3.

Let \mathcal{A} be a proper learning algorithm for \mathcal{H} , let $n \geq d$, and assume, for the eventual purpose of obtaining a contradiction, that $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq f(d)$ for all \mathcal{D} and, for all $s \in \mathcal{Z}^n$, $\mathbb{E}\hat{R}_s(\mathcal{A}_n(s)) \leq cd/n$ if there exists $h \in \mathcal{H}$ such that $\hat{R}_s(h) = 0$. Pick a realizable distribution \mathcal{D} . It follows from the above assumption and Eq. (4.1) that $\mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n))] \leq 2cd/n + 3f(d)/n = (3f(d) + 2cd)/n$. By Markov’s inequality, $\mathbb{P}(R_{\mathcal{D}}(\mathcal{A}_n(S_n)) \geq \epsilon) \leq \frac{1}{n\epsilon}(3f(d) + 2cd)$. It follows that the sample complexity of proper learning \mathcal{H} satisfies

$$\mathcal{M}_{\text{prop}}^{\mathcal{H}}(\epsilon, \delta) \leq \frac{1}{\epsilon\delta}(3f(d) + 2cd). \quad (\text{C.3})$$

Now, fix $\delta \in (0, 1/100)$ and fix a convergent sequence of $\epsilon_i \downarrow 0$. There exists J such

that, for all $i \geq J$,

$$\frac{1}{\tilde{c}\delta}(3f(d) + 2cd) < d \operatorname{Log} \frac{1}{\epsilon_i} + \operatorname{Log} \frac{1}{\delta}, \quad (\text{C.4})$$

for \tilde{c} as in Theorem 4.4.3. Combining Eq. (C.4) with Eq. (C.3), $\mathcal{M}_{\text{prop}}^{\mathcal{H}}(\epsilon_i, \delta) < \frac{\tilde{c}}{\epsilon_i}(d \operatorname{Log} \frac{1}{\epsilon_i} + \operatorname{Log} \frac{1}{\delta})$ for $i \geq J$. Simultaneously, from Theorem 4.4.3, it follows that $\mathcal{M}_{\text{prop}}^{\mathcal{H}}(\epsilon_i, \delta) \geq \frac{\tilde{c}}{\epsilon_i}(d \operatorname{Log} \frac{1}{\epsilon_i} + \operatorname{Log} \frac{1}{\delta})$, a contradiction.

C.4 Proof of Theorem 4.4.6

Let $n \geq \mathfrak{s}$. First we begin with two definitions: letting S_n be the random element in $(\mathcal{X} \times \mathcal{Y})^n$ representing our training sample, the *empirical teaching dimension* [Han07; EW10], denoted ETD_n , is size of the smallest subset of S that produces the same version space, i.e.,

$$\text{ETD}_n = \min\{|S'| : S' \subseteq S_n, V_{\mathcal{H}}[S'] = V_{\mathcal{H}}[S_n]\}. \quad (\text{C.5})$$

An important fact about the empirical teaching dimension is that ETD_n is bounded by star number almost surely [Han16]. An *empirical teaching set* is any subset of S_n that achieves the minimum in the definition of the empirical teaching dimension. Consider a realizable training set $S_n \in (\mathcal{X} \times \mathcal{Y})^n$. Let S' denote an empirical teaching set of S_n . Let $\hat{S} \subseteq S_n \setminus S'$. By the definition of the version space we have $V_{\mathcal{H}}[S_n] \subseteq V_{\mathcal{H}}[S_n \setminus \hat{S}] \subseteq V_{\mathcal{H}}[S']$. Also, by the definition of S' , we have $V_{\mathcal{H}}[S_n] = V_{\mathcal{H}}[S']$; therefore, $V_{\mathcal{H}}[S_n] = V_{\mathcal{H}}[S_n \setminus \hat{S}]$. This argument shows that the version space is *stable*, in the sense that removing any point that is not in S' does not alter the version space. Let $\text{ETS}(S_n)$ denote the empirical teaching set S_n .

We also need the definition of the *region of disagreement* denoted by $\text{DIS}(V_{\mathcal{H}}[S_n]) = \{x \in \mathcal{X} \mid \exists h, h' \in V_{\mathcal{H}}[S_n] \text{ such that } h(x) \neq h'(x)\}$.

Following the same line of reasoning as in Eq. (4.4) we obtain $I(V_{\mathcal{H}}[Z_U]; U|Z) \leq n \log 2 - \sum_{i=1}^n H(U_i|U_{-i}, V_{\mathcal{H}}[Z_U], Z)$. Since the order of the training set does not change the version space we get $\sum_{i=1}^n H(U_i|U_{-i}, V_{\mathcal{H}}[Z_U], Z) = nH(U_1|U_{-1}, V_{\mathcal{H}}[Z_U], Z)$. Fix $i \in [n]$, and define $U_{i \rightarrow b} \triangleq (U_1, \dots, U_{i-1}, b, U_{i+1}, \dots, U_n)$ for $b \in \{0, 1\}$. Using this notation we can define training sets $S_{i \rightarrow b} = Z_{U_{i \rightarrow b}}$ for $b \in \{0, 1\}$ and $i \in [n]$. Let $\mathcal{F}_1 = \sigma(U_{-1}, V_{\mathcal{H}}[Z_U], Z)$. Then, we have

$$H(U_1|U_{-1}, V_{\mathcal{H}}[Z_U], Z) \geq \mathbb{E}[H^{\mathcal{F}_1}(U_1) \mathbf{1}[V_{\mathcal{H}}[S_{1 \rightarrow 0}] = V_{\mathcal{H}}[S_{1 \rightarrow 1}]]]. \quad (\text{C.6})$$

Using the same techniques as in the proof Theorem 4.3.4, we can show that on the event $\mathbb{1}[V_{\mathcal{H}}[S_{1 \rightarrow 0}] = V_{\mathcal{H}}[S_{1 \rightarrow 1}]]$, $H^{\mathcal{F}_1}(U_1) = \log(2)$. Thus, $H(U_1|U_{-1}, V_{\mathcal{H}}[Z_U], Z) \geq \mathbb{P}(V_{\mathcal{H}}[S_{1 \rightarrow 0}] = V_{\mathcal{H}}[S_{1 \rightarrow 1}]) \log 2$.

For $i \in [n]$, define the training set $S_{-i} \triangleq \{Z_{U_1}, \dots, Z_{U_{i-1}}, Z_{U_{i+1}}, \dots, Z_{U_n}\}$. Recall that $Z_{i,j} = (X_{i,j}, Y_{i,j})$. We claim that

$$V_{\mathcal{H}}[S_{1 \rightarrow 0}] \neq V_{\mathcal{H}}[S_{1 \rightarrow 1}] \Rightarrow (X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])) \vee (X_{1,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])). \quad (\text{C.7})$$

We prove this claim by contraposition. Given that $(X_{0,1} \notin \text{DIS}(V_{\mathcal{H}}[S_{-1}])) \wedge (X_{1,1} \notin \text{DIS}(V_{\mathcal{H}}[S_{-1}]))$, we have that the concepts in $\text{DIS}(V_{\mathcal{H}}[S_{-1}])$ agree for prediction of $X_{0,1}$ and $X_{1,1}$. As \mathcal{D} is a realizable distribution, we conclude $V_{\mathcal{H}}[S_{1 \rightarrow 0}] = V_{\mathcal{H}}[S_{-1}] = V_{\mathcal{H}}[S_{1 \rightarrow 1}]$.

In the next step, we provide an upper bound on $\mathbb{P}((X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])) \vee (X_{1,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])))$ as follows

$$\begin{aligned} & \mathbb{P}((X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])) \vee (X_{1,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}]))) \\ & \leq \mathbb{P}(X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}]))) + \mathbb{P}(X_{1,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}]))) \\ & = 2\mathbb{P}(X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}]))). \end{aligned} \quad (\text{C.8})$$

Here, we have used the union bound and the points in Z are i.i.d.. Then, we can write

$$\begin{aligned} \mathbb{P}(X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])) &= \mathbb{E}[\mathbb{1}[X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])]] \\ &= \mathbb{E}[\mathbb{1}[X_1 \in \text{DIS}(V_{\mathcal{H}}[\{Z_2, \dots, Z_n\}])]]. \end{aligned} \quad (\text{C.9})$$

The last step follows from U and Z are independent, and the points in Z are i.i.d.. Therefore, in the last step we consider the expectation over $(Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$. Note that $Z_i = (X_i, Y_i)$. By the exchangeability of the points in (Z_1, \dots, Z_n) we have

$$\mathbb{E}[\mathbb{1}[X_1 \in \text{DIS}(V_{\mathcal{H}}[\{Z_2, \dots, Z_n\}])]] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbb{1}[X_i \in \text{DIS}(V_{\mathcal{H}}[\{Z_1, \dots, Z_n\} \setminus \{Z_i\}])]]. \quad (\text{C.10})$$

Then, we claim that given $X_i \in \text{DIS}(V_{\mathcal{H}}[\{Z_1, \dots, Z_n\} \setminus \{Z_i\}])$ then $X_i \in S'$ where S' is *any* teaching set of $\{Z_1, \dots, Z_n\}$. We can easily prove this claim by contradiction

and the stability of the version space shown in the beginning of this section. Therefore,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbb{1}[X_i \in \text{DIS}(V_{\mathcal{H}}[\{Z_1, \dots, Z_n\} \setminus \{Z_i\}])]] &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbb{1}[X_i \in S']] \\
&= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \mathbb{1}[X_i \in S'] \right] \\
&\leq \frac{\mathfrak{s}}{n}. \tag{C.11}
\end{aligned}$$

Here, we have used the linearity of the expectation, and the last step follows from the fact that the cardinality of S' is at most \mathfrak{s} almost surely. By Eq. (C.8)-Eq. (C.11), we have $\mathbb{P}((X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])) \vee (X_{1,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}]))) \leq 2\mathfrak{s}/n$. Then, by Eq. (C.7) we have

$$\begin{aligned}
H(U_1|U_{-1}, V_{\mathcal{H}}[Z_U], Z) &\geq \mathbb{P}(V_{\mathcal{H}}[S_{1 \rightarrow 0}] = V_{\mathcal{H}}[S_{1 \rightarrow 1}]) \log 2 \\
&\geq [1 - \mathbb{P}((X_{0,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])) \vee (X_{1,1} \in \text{DIS}(V_{\mathcal{H}}[S_{-1}])))] \log 2 \\
&\geq (1 - \frac{2\mathfrak{s}}{n}) \log 2. \tag{C.12}
\end{aligned}$$

Finally, combining this results and $I(V_{\mathcal{H}}[Z_U]; U|Z) \leq n \log 2 - nH(U_1|U_{-1}, V_{\mathcal{H}}[Z_U], Z)$, we obtain $I(V_{\mathcal{H}}[Z_U]; U|Z) \leq 2\mathfrak{s} \log 2$ which was to be shown.

C.5 Proof of Theorem 4.4.8

We begin the proof by a lemma, which can be seen as the generalization of the well-known coupon collector's problem [BHMZ20, Lem. 19].

Lemma C.5.1. *Let $M, m \in \mathbb{N}$ and $1 \leq m \leq M$. Assume we take k samples X_1, \dots, X_k uniformly at random with replacement from $[M]$. Then $\mathbb{P}(|[M] \setminus \{X_1, \dots, X_k\}| \geq m) \geq 1/2$ if $k \leq (M/2) \log \frac{M}{m}$.*

Note that $|[M] \setminus \{X_1, \dots, X_k\}|$ is the number of unseen elements from $[M]$ after observing samples X_1, \dots, X_k .

Consider a concept class \mathcal{H} over the input space \mathcal{X} whose star number is \mathfrak{s} . Let $M = \min\{n, \mathfrak{s}\}$. By the definition of star number, there exists $(x_1, \dots, x_M) \in \mathcal{X}^M$, such that $((x_1, y_1), \dots, (x_M, y_M))$ is realizable by $h_0^* \in \mathcal{H}$, and every neighbour of this sequence is also realizable by a classifier in \mathcal{H} . Let $h_{x_i} \in \mathcal{H}$ be any classifier such that $\{j \in [M] | h_{x_i}(x_j) \neq y_j\} = \{i\}$.

For the case $M = 1$, our lower bound is simply $\min\{n, \mathfrak{s}\} - 1 = M - 1 = 0$ which is trivial since the mutual information is non-negative. Therefore, in the rest of the proof we assume $M \geq 2$.

For the case $M \geq 2$, consider the following distribution on the input space

$$\mathcal{D}_X(x_1) = 1 - \frac{M-1}{n} \text{ and } \mathcal{D}_X(x_i) = \frac{1}{n}, \text{ for } i \in \{2, \dots, M\}.$$

Also let the target function (labelling function) be h_0^* . Let Z , U , and S be defined as usual, based on a sample from \mathcal{D}_X labeled by h_0^* . Since \mathcal{D}_X has zero measure on $\mathcal{X} \setminus \{x_1, \dots, x_M\}$, we can, without any loss of generality, assume that $\mathcal{H} = \{h_0^*, h_{x_1}, \dots, h_{x_M}\}$, as every other classifier is equivalent to one of these \mathcal{D}_X -almost everywhere. When we release the version space, we can agree in advance that elements stand for their equivalence classes, which does not affect the information content.

Let $X = (X_{i,j})_{i \in \{0,1\}, j \in [n]}$ denote the inputs observed in the supersample Z . Define X_U as the sequence of the inputs observed in Z_U . Similarly, let $X_{\bar{U}}$ denote the sequence of inputs observed in the ‘‘ghost sample’’ $Z_{\bar{U}}$, where \bar{U} denotes the sequence U but with every entry flipped. In the following, we write $x \in X_U$ to mean that there exists $i \in [n]$ such that $X_{U_i,i} = x$.

We claim that $V_{\mathcal{H}}[Z_U] = \{h_{x_i} | i \in [M], x_i \notin X_U\} \cup \{h_0^*\}$.¹ To see this, note that, if $x_i \notin X_U$, then $h_{x_i} = h_0^*$ on S and so both are ERMs. In the other direction, if $x_i \in X_U$, then h_{x_i} makes at least one mistake, and so is not an ERM. Thus, if $V_{\mathcal{H}}[Z_U]$ has a non-empty intersection with $\{h_{X_{0,i}}, h_{X_{1,i}}\}$, we can perfectly recover U_i from $V_{\mathcal{H}}[Z_U]$, since in each column of the supersample Z , only one point is selected for the training set. We use this observation to lower bound the conditional entropy of U given Z and the version space.

To that end, let $J = \{j \in [n] | V_{\mathcal{H}}[Z_U] \cap \{h_{X_{0,j}}, h_{X_{1,j}}\} \neq \emptyset\}$ be the set that contains the index of every column j for which we can perfectly recover U_j from the version space. Note that we can represent J in an equivalent form as follows. For two sequences $v, w \in \{x_1, \dots, x_M\}^n$, define $K(v, w) = \{j \in [n] | w_j \neq v_i \text{ for all } i \in [n]\}$. By the definition of the version space, $x_i \notin X_U$ is equivalent to $h_{x_i} \in V_{\mathcal{H}}[Z_U]$. Let $j \in [n]$, then

$$j \in J \Leftrightarrow V_{\mathcal{H}}[Z_U] \cap \{h_{X_{1-\bar{U}_j,j}}\} = \{h_{X_{1-\bar{U}_j,j}}\} \Leftrightarrow X_{1-\bar{U}_j,j} \notin X_U \Leftrightarrow j \in K(X_U, X_{\bar{U}}).$$

Therefore, $J = K(X_U, X_{\bar{U}})$.

¹Formally, this statement holds almost surely. We will skip such declarations for the remainder of the proof.

By the definition of the mutual information, and the fact that U and Z are independent, we have

$$I(\mathcal{V}_{\mathcal{H}}[Z_U]; U|Z) = n \log(2) - H(U|Z, \mathcal{V}_{\mathcal{H}}[Z_U]). \quad (\text{C.13})$$

Then

$$H(U|Z, \mathcal{V}_{\mathcal{H}}[Z_U]) = H(U|Z, \mathcal{V}_{\mathcal{H}}[Z_U], J) \quad (\text{C.14})$$

$$= H(U_J, U_{J^c}|Z, \mathcal{V}_{\mathcal{H}}[Z_U], J) \quad (\text{C.15})$$

$$= H(U_{J^c}|Z, \mathcal{V}_{\mathcal{H}}[Z_U], J) \quad (\text{C.16})$$

$$\leq \mathbb{E}[n - |J|] \log(2), \quad (\text{C.17})$$

where Eq. (C.14) follows from J being known from Z and $\mathcal{V}_{\mathcal{H}}[Z_U]$; Eq. (C.16) follows from U_J being known from J , Z , $\mathcal{V}_{\mathcal{H}}[Z_U]$; and Eq. (C.17) follows from the cardinality of the support of the distribution of U_{J^c} being no more than $2^{(n-|J|)}$. Therefore, by Eq. (C.13) and Eq. (C.17), we have

$$I(\mathcal{V}_{\mathcal{H}}[Z_U]; U|Z) \geq \mathbb{E}[|J|] \log 2. \quad (\text{C.18})$$

In the next step of the proof, we lower bound $\mathbb{E}[|J|]$. Because U and Z are independent and Z is an i.i.d. array, X_U and $X_{\bar{U}}$ are independent and identically distributed sequences of i.i.d. elements with common distribution $\mathcal{D}_{\mathcal{X}}$. Thus, $\mathbb{E}[|J|] = \mathbb{E}[|K(X, \bar{X})|]$, where X and \bar{X} are i.i.d. copies of X_U (equivalently, $X_{\bar{U}}$).

Let \hat{M} be the number of elements of $\{x_2, \dots, x_M\}$ not appearing in X , i.e., $\hat{M} = |\{i \in \{2, \dots, M\} | x_i \neq X_j \text{ for all } j \in [n]\}|$. Conditional on X , if $x_1 \in X$, then $|K(X, \bar{X})|$ is a Binomial random variable with n trials, each succeeding with probability $\frac{\hat{M}}{n}$. If $x_1 \notin X$, then $|K(X, \bar{X})|$ is a Binomial random variable with n trials, each succeeding with probability $\frac{\hat{M}}{n} + 1 - \frac{M-1}{n} \geq \frac{\hat{M}}{n}$. Therefore,

$$\mathbb{E}[|J|] = \mathbb{E}[|K(X, \bar{X})|] \quad (\text{C.19})$$

$$= \mathbb{E}[\mathbb{E}^X[|K(X, \bar{X})|]] \quad (\text{C.20})$$

$$\geq \mathbb{E}[\hat{M}]. \quad (\text{C.21})$$

Here, Eq. (C.20) follows from the tower rule and Eq. (C.21) follows from the fact that the mean of the binomial distribution with n trials, each succeeding with probability p , is np .

Fix $\beta = \exp(-3)$ and $m = \beta(M-1)$. Let \hat{N} denote the number of samples in X

falling in $\{x_2, \dots, x_M\}$. Then, using the tower rule and Markov's inequality we have

$$\mathbb{E}[\hat{M}] = \mathbb{E}[\mathbb{E}^{\hat{N}}[\hat{M}]] \quad (\text{C.22})$$

$$\geq \mathbb{E}[m\mathbb{P}^{\hat{N}}[\hat{M} \geq m]] \quad (\text{C.23})$$

$$\geq \mathbb{E}[m\mathbb{P}^{\hat{N}}[\hat{M} \geq m]\mathbb{1}[\hat{N} \leq ((M-1)/2)\log\beta^{-1}]] \quad (\text{C.24})$$

Note that, conditional on \hat{N} , the \hat{N} samples falling in $\{x_2, \dots, x_M\}$ are conditionally independent, with conditional distribution uniform on this set. Using this fact, from Lemma C.5.1 we obtain

$$\mathbb{E}[m\mathbb{P}^{\hat{N}}[\hat{M} \geq m]\mathbb{1}[\hat{N} \leq ((M-1)/2)\log\beta^{-1}]] \geq \frac{1}{2}m\mathbb{P}(\hat{N} \leq ((M-1)/2)\log\beta^{-1}). \quad (\text{C.25})$$

We have $\mathbb{E}[\hat{N}] = M-1$ and, by Markov's inequality, $\mathbb{P}(\hat{N} \leq \frac{M-1}{2}\log\beta^{-1}) \geq 1 - 2/\log(\beta^{-1}) = 1/3$. By combining Eq. (C.18), Eq. (C.21), Eq. (C.25), and $\mathbb{P}(\hat{N} \leq \frac{M-1}{2}\log\beta^{-1}) \geq 1/3$ we get

$$\begin{aligned} I(\mathbb{V}_{\mathcal{H}}[Z_U]; U|Z) &\geq \frac{\beta \log 2}{6}(M-1) \\ &= \Omega(\min\{n, \mathfrak{s}\} - 1), \end{aligned}$$

which was to be shown.

C.6 Proof of Theorem 4.4.9

By the well-ordering theorem [Zer08], there exists a binary relation \ll on \mathcal{H} that is transitive, total, antisymmetric, and well-ordered. In particular, the well-ordered property implies that every nonempty subset \mathcal{H} has the least element. The proposed learning algorithm is given by

$$\mathcal{A}_n(S_n) = \text{LE}(\mathbb{V}_{\mathcal{H}}[S_n]),$$

where LE of a nonempty set denotes its least element with respect to \ll .² Note that \mathcal{A}_n is *deterministic*, *consistent*, and *permutation-invariant*, i.e., the order of the points in S_n does not impact the output.

²In some cases this classifier may not be measurable. We will assume it is. To avoid measure-theoretic issues, one may assume \mathcal{X} is countably infinite or finite, or design a well-ordering by hand to guarantee measurability if possible.

Let $W = \mathcal{A}_n(Z_U)$. By the definition of the mutual information and Lemma 4.2.1, we obtain

$$\begin{aligned} \text{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= H(U|Z) - H(U|W, Z) \\ &\leq n \log(2) - \sum_{i=1}^n H(U_i|U_{-i}, Z, W). \end{aligned} \quad (\text{C.26})$$

The last step follows since U is independent of Z and the independence of indices of U and Lemma 4.2.1. By the permutation-invariance of the algorithm, we have $\sum_{i=1}^n H(U_i|U_{-i}, Z, W) = nH(U_1|U_{-1}, Z, W)$. Fix $i \in [n]$, and define $U_{i \rightarrow j} \triangleq (U_1, \dots, U_{i-1}, j, U_{i+1}, \dots, U_n)$ for $j \in \{0, 1\}$. Using this notation we can define training set $S_{i \rightarrow j} = Z_{U_{i \rightarrow j}}$ for $j \in \{0, 1\}$ and $U_{-i} \in \{0, 1\}^{n-1}$. Let $\mathcal{F}_1 = \sigma(W, U_{-1}, Z)$. Define the \mathcal{F}_1 -measurable event $\mathcal{E} = \{\mathcal{A}_n(S_{1 \rightarrow 0}) = \mathcal{A}_n(S_{1 \rightarrow 1})\}$. Then, we can write $H(U_1|U_{-1}, Z, W) = \mathbb{E}[H^{\mathcal{F}_1}(U_1)(\mathbb{1}[\mathcal{E}] + \mathbb{1}[\mathcal{E}^c])]$. We claim the following facts:

1. On the event \mathcal{E}^c , $H^{\mathcal{F}_1}(U_1) = 0$ a.s..
2. On the event \mathcal{E} , $H^{\mathcal{F}_1}(U_1) = \log 2$ a.s..

The reason is that since the algorithm is deterministic, on the event \mathcal{E}^c we can perfectly recover U_1 since W is either $\mathcal{A}_n(S_{1 \rightarrow 0})$ or $\mathcal{A}_n(S_{1 \rightarrow 1})$ which shows that $H^{\mathcal{F}_1}(U_1) = 0$. Then, on the event \mathcal{E} , using the Bayes rule we can show that $H^{\mathcal{F}_1}(U_1) = \log 2$. Therefore, we have $H(U_1|U_{-1}, Z, W) = \mathbb{E}[\mathbb{1}[\mathcal{E}]] \log(2)$. We can further lower bound

$$H(U_1|U_{-1}, Z, W) \geq \mathbb{E}[\mathbb{1}[\mathcal{E}]\mathbb{1}[\ell(\mathcal{A}_n(S_{1 \rightarrow 0}), Z_{1,1}) = 0 \wedge \ell(\mathcal{A}_n(S_{1 \rightarrow 1}), Z_{0,1}) = 0]] \log(2). \quad (\text{C.27})$$

Next, we claim that on the event $\{\ell(\mathcal{A}_n(S_{1 \rightarrow 0}), Z_{1,1}) = 0\} \wedge \{\ell(\mathcal{A}_n(S_{1 \rightarrow 1}), Z_{0,1}) = 0\}$, we have $\mathcal{A}_n(S_{1 \rightarrow 0}) = \mathcal{A}_n(S_{1 \rightarrow 1})$ a.s.. This claim can be proved by contradiction. Assume $\mathcal{A}_n(S_{1 \rightarrow 0}) \neq \mathcal{A}_n(S_{1 \rightarrow 1})$. Consider the version space $V_{\mathcal{H}}[S_{1 \rightarrow 0} \cup S_{1 \rightarrow 1}]$. By the assumptions $\ell(\mathcal{A}_n(S_{1 \rightarrow 0}), Z_{1,1}) = 0$ and $\ell(\mathcal{A}_n(S_{1 \rightarrow 1}), Z_{0,1}) = 0$, we have $\mathcal{A}_n(S_{1 \rightarrow 0}) \in V_{\mathcal{H}}[S_{1 \rightarrow 0} \cup S_{1 \rightarrow 1}]$ and $\mathcal{A}_n(S_{1 \rightarrow 1}) \in V_{\mathcal{H}}[S_{1 \rightarrow 0} \cup S_{1 \rightarrow 1}]$. Also, it is immediate to see that $V_{\mathcal{H}}[S_{1 \rightarrow 0} \cup S_{1 \rightarrow 1}] \subseteq V_{\mathcal{H}}[S_{1 \rightarrow 0}]$ and $V_{\mathcal{H}}[S_{1 \rightarrow 0} \cup S_{1 \rightarrow 1}] \subseteq V_{\mathcal{H}}[S_{1 \rightarrow 1}]$. Therefore, we have $\mathcal{A}_n(S_{1 \rightarrow 0}) \in V_{\mathcal{H}}[S_{1 \rightarrow 1}]$ and $\mathcal{A}_n(S_{1 \rightarrow 1}) \in V_{\mathcal{H}}[S_{1 \rightarrow 0}]$. By the definition of the algorithm, we choose the least hypothesis from the version space. Thus, $\mathcal{A}_n(S_{1 \rightarrow 0}) \ll \mathcal{A}_n(S_{1 \rightarrow 1})$ since $\mathcal{A}_n(S_{1 \rightarrow 1}) \in V_{\mathcal{H}}[S_{1 \rightarrow 0}]$. Similarly, we can show $\mathcal{A}_n(S_{1 \rightarrow 1}) \ll \mathcal{A}_n(S_{1 \rightarrow 0})$. Therefore, considering $\mathcal{A}_n(S_{1 \rightarrow 1}) \ll \mathcal{A}_n(S_{1 \rightarrow 0})$ and $\mathcal{A}_n(S_{1 \rightarrow 0}) \ll \mathcal{A}_n(S_{1 \rightarrow 1})$, we conclude that the assumption $\mathcal{A}_n(S_{1 \rightarrow 0}) \neq \mathcal{A}_n(S_{1 \rightarrow 1})$ is false, a contradiction.

Having proved that $\mathcal{A}_n(S_{1 \rightarrow 0}) = \mathcal{A}_n(S_{1 \rightarrow 1})$ on the event $\{\ell(\mathcal{A}_n(S_{1 \rightarrow 0}), Z_{1,1}) = 0\} \wedge \{\ell(\mathcal{A}_n(S_{1 \rightarrow 1}), Z_{0,1}) = 0\}$, we can further simplify Eq. (C.27) as

$$\begin{aligned} H(U_1|U_{-1}, Z, W) &\geq \mathbb{E}[\mathbb{1}[\ell(\mathcal{A}_n(S_{1 \rightarrow 1}), Z_{0,1}) = 0]\mathbb{1}[\ell(\mathcal{A}_n(S_{1 \rightarrow 0}), Z_{1,1}) = 0]] \log 2 \\ &\geq (1 - \mathbb{E}[\mathbb{1}[\ell(\mathcal{A}_n(S_{1 \rightarrow 1}), Z_{0,1}) = 1]] - \mathbb{E}[\mathbb{1}[\ell(\mathcal{A}_n(S_{1 \rightarrow 1}), Z_{0,1}) = 1]]) \log 2 \\ &= (1 - 2\mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n))]) \log 2. \end{aligned}$$

The last step follows since the elements of $S_{1 \rightarrow 0}$ and $S_{1 \rightarrow 1}$ are i.i.d.. By plugging this lower bound into Eq. (C.26), we obtain

$$\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 2n \log(2) \mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n))]. \quad (\text{C.28})$$

Then, we use the result from Corollary 12 of [Han16] which states the following bound holds uniformly for the expected risk of *all* consistent and proper learners for learning a concept class with VC dimension d and star number \mathfrak{s} :

$$\mathbb{E}[R_{\mathcal{D}}(\mathcal{A}_n(S_n))] = O\left(\frac{d}{n} \log\left(\frac{\min\{\mathfrak{s}, n\}}{d}\right)\right).$$

Finally, the stated result follows by combining this bound with Eq. (C.28).

C.7 Proof of Theorem 4.5.1

Let $L = (L_1, \dots, L_n)$ where $L_i \in \{0, 1\}^2$ denotes the vector at Column i of the loss array L . By the definition of mutual information, we have $I(L; U|Z) \leq H(L|Z)$. Since conditioning decreases the entropy, we have $H(L|Z) \leq H(L)$. Finally, by the chain rule, $H(L) \leq \sum_{i=1}^n H(L_i)$.

Note that L_i takes values in $\{[1, 0]^\top, [0, 0]^\top, [0, 1]^\top\}$, as it is assumed that \mathcal{A}_n is consistent and, by construction, at each column, one of the points is selected for the training set. If $U_i = 0$, then $Z_{0,i}$ is in the training set and so, because the algorithm is consistent, $L_{0,i} = 0$ a.s. Therefore,

$$\begin{aligned} \mathbb{P}[L_i = [1, 0]^\top] &= \mathbb{E}[\mathbb{P}^U[L_i = [1, 0]^\top]] \\ &= \mathbb{E}[\mathbb{1}[U_i = 0]\mathbb{P}^U[L_i = [1, 0]^\top]] + \mathbb{E}[\mathbb{1}[U_i = 1]\mathbb{P}^U[L_i = [1, 0]^\top]] \\ &= \mathbb{E}[\mathbb{1}[U_i = 1]\mathbb{P}^U[L_i = [1, 0]^\top]] \end{aligned}$$

Conditioning on event $U_i = 1$, we have $L_i = [1, 0]^\top = \ell(\mathcal{A}_n(Z_U), Z_{0,i}) = 1$ and therefore

$$\mathbb{P}^U[L_i = [1, 0]^\top] = \mathbb{P}^U[\ell(\mathcal{A}_n(Z_U), Z_{0,i}) = 1] = R_{\mathcal{D}}(\mathcal{A}_n),$$

where the final equality follows from the definition of the expected risk and the fact that $Z_{0,i}$ is not in Z_U . Therefore, $\mathbb{P}[L_i = [1, 0]^\top] = \mathbb{E}[\mathbf{1}[U_i = 1]]R_{\mathcal{D}}(\mathcal{A}_n) = (1/2)R_{\mathcal{D}}(\mathcal{A}_n)$. The same idea establishes that $\mathbb{P}[L_i = [0, 1]^\top] = (1/2)R_{\mathcal{D}}(\mathcal{A}_n)$.

Therefore, by the definition of the entropy

$$\begin{aligned} H(L_i) &= -(1 - R_{\mathcal{D}}(\mathcal{A}_n)) \log(1 - R_{\mathcal{D}}(\mathcal{A}_n)) - R_{\mathcal{D}}(\mathcal{A}_n) \log\left(\frac{R_{\mathcal{D}}(\mathcal{A}_n)}{2}\right) \\ &= H_b(R_{\mathcal{D}}(\mathcal{A}_n)) + R_{\mathcal{D}}(\mathcal{A}_n) \log(2). \end{aligned}$$

The stated results follows from $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) \leq \sum_{i=1}^n H(L_i)$.

C.8 Proof of Theorem 4.5.6

We begin by presenting a theorem that will be used later to prove Theorem 4.5.6. This theorem, which might be of independent interest, shows the *average leave-one-error over supersample* Z can be used to upper bound $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))$.

Lets begin by introducing some notations. For $n \in \mathbb{N}$, let $[2n]_{n+1}$ denote the set of all size- $n + 1$ subsets of $[2n]$ and let $H_b(\cdot)$ denote the binary entropy function.

Theorem C.8.1. *Let \mathcal{A} denote a consistent and symmetric algorithm. Let $P_e : \mathcal{Z}^n \times \mathcal{Z} \rightarrow [0, 1]$. For $s = ((x_1, y_1), \dots, (x_n, y_n)) \in \mathcal{Z}^n$ and $(x, y) \in \mathcal{Z}$, $P_e(s; (x, y))$ denotes the probability of error \mathcal{A}_n for predicting the label x where the randomness is over the internal randomness in \mathcal{A} . Then, for every distribution \mathcal{D} , we have*

$$\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) \leq n \mathbb{E} \left[H_b(\kappa(\tilde{Z})) + \kappa(\tilde{Z}) \log 2 - \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} H_b(P_e(\tilde{Z}_{J-\{j\}}; \tilde{Z}_j)) \right] \quad (\text{C.29})$$

where $\tilde{Z} = (Z_1, \dots, Z_{2n}) \sim \mathcal{D}^{\otimes(2n)}$ and $\kappa(\tilde{Z}) = \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} P_e(\tilde{Z}_{J-\{j\}}; \tilde{Z}_j)$ which takes values in $[0, 1]$ almost surely. Note, in Eq. (C.29), the outer expectation is over \tilde{Z} .

The proof of Theorem C.8.1 is deferred to Appendix C.8.2.

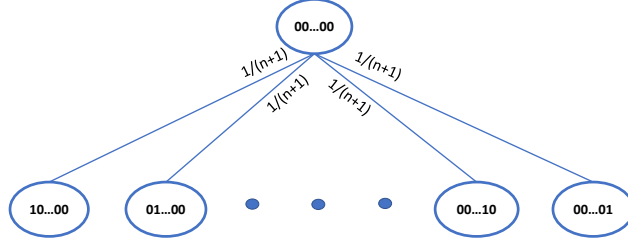


Figure C.1: One-inclusion graph of point functions for a set of distinct points.

C.8.1 Proof of Theorem 4.5.6

First of all, note that given *distinct* points $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}$, the vertex set of the one-inclusion graph is given by

$$\{(h(x_1), \dots, h(x_{n+1})) | h \in \mathcal{H}\} = \{(a_1, \dots, a_{n+1}) | a_i \in \{0, 1\} \text{ for all } i \in [n+1], \sum_{i \in [n+1]} a_i \leq 1\}.$$

As an example Fig. C.1 illustrates the structure of the one-inclusion graph for a sequence of *distinct* points. The prediction rule is as follows. Given a training set $((x_1, y_1), \dots, (x_n, y_n))$ and test example x , we have two cases. If $y_i = 0$ for all $i \in [n]$, then the label of x is predicted to be 0 with probability $\frac{n}{n+1}$, and it is predicted to be one with probability $\frac{1}{n+1}$. The second case is that there exists a point with label 1. Denote its index by $i^* \in [n]$. In this case, the target function is known to be $\mathbb{1}[x = x_{i^*}]$.

To upper bound eCMI for this prediction rule, we consider the upper bound provided in Theorem C.8.1. For a fixed supersample $\tilde{Z} = ((x_1, y_1), \dots, (x_{2n}, y_{2n}))$, we will consider two cases:

Case I: For all $i \in [2n]$ we have $y_i = 0$. Let $J \in [2n]_{n+1}$ and let $\deg(\tilde{Z}_J)$ denote the degree of the vertex $(0, 0, \dots, 0, 0)$ in the one-inclusion graph \tilde{Z}_J . Note that if there is no repetition in \tilde{Z}_J then, the degree is $n+1$. If repetition exists, it is less than $n+1$. Then, according to the prediction rule we have

$$\begin{aligned} \kappa(\tilde{Z}) &= \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} P_e(\tilde{Z}_{J-\{j\}}, \tilde{Z}_j) \\ &= \frac{1}{n+1} \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{\deg(\tilde{Z}_J)}{n+1}. \end{aligned} \tag{C.30}$$

Also,

$$\mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} H_b(P_e(\tilde{Z}_{J-\{j\}}, \tilde{Z}_j)) = H_b\left(\frac{1}{n+1}\right) \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{\deg(\tilde{Z}_J)}{n+1}. \quad (\text{C.31})$$

let $\alpha = \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{\deg(\tilde{Z}_J)}{n+1}$. If $\alpha = 0$, then $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) = 0 = O(1)$. Therefore, assume $\alpha > 0$, then we can use Eq. (C.29), Eq. (C.30), and Eq. (C.31) to obtain

$$\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) \leq n \mathbb{E} \left[H_b\left(\frac{\alpha}{n+1}\right) + \frac{\alpha \log 2}{n+1} - \alpha H_b\left(\frac{1}{n+1}\right) \right] \quad (\text{C.32})$$

$$\begin{aligned} &\leq n \mathbb{E} \left[-\frac{\alpha}{n+1} \log \frac{\alpha}{n+1} - \left(1 - \frac{\alpha}{n+1}\right) \log \left(1 - \frac{\alpha}{n+1}\right) + \frac{\alpha \log 2}{n+1} + \right. \\ &\quad \left. \frac{\alpha}{n+1} \log \left(\frac{1}{n+1}\right) + \alpha \left(1 - \frac{1}{n+1}\right) \log \left(1 - \frac{1}{n+1}\right) \right] \quad (\text{C.33}) \end{aligned}$$

$$\leq n \mathbb{E} \left[-\frac{\alpha}{n+1} \log(\alpha) + \frac{2\alpha \log 2}{n+1} \right] \quad (\text{C.34})$$

$$\leq n \left[\frac{2 \log 2}{n+1} + \frac{\exp(-1)}{n+1} \right] \quad (\text{C.35})$$

$$= O(1). \quad (\text{C.36})$$

Eq. (C.34) is obtained by removing the negative terms and the inequality $-(1-x) \log(1-x) \leq x \log 2$ for $x \in [0, 1]$. Eq. (C.35) follows from $\max_{x \in [0, 1]} -x \log(x) = \exp(-1)$ and $\alpha \leq 1$.

Case II: There exists $i \in [2n]$ such that $y_i = 1$. Let $m = |\{i \in [2n] | y_i = 1\}|$, $\beta = \frac{\binom{2n-m}{n+1}}{\binom{2n}{n+1}}$, and $\alpha = \mathbb{E}_J \left[\frac{\deg(Z_J)}{n+1} \mid \text{no point with label 1 in } \tilde{Z}_J \right]$. Then, we have

$$\begin{aligned} \kappa(\tilde{Z}) &= \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} P_e(\tilde{Z}_{J-\{j\}}, \tilde{Z}_j) \\ &= \beta \mathbb{E}_J \left[\frac{1}{n+1} \sum_{j \in J} P_e(\tilde{Z}_{J-\{j\}}, \tilde{Z}_j) \mid \text{no point with label 1 in } \tilde{Z}_J \right] \\ &\quad + (1-\beta) \mathbb{E}_J \left[\frac{1}{n+1} \sum_{j \in J} P_e(\tilde{Z}_{J-\{j\}}, \tilde{Z}_j) \mid \text{there exists points with label 1 in } \tilde{Z}_J \right] \\ &\leq \alpha \beta \frac{1}{n+1} + (1-\beta) \frac{1}{(n+1)^2}, \quad (\text{C.37}) \end{aligned}$$

where in the last step we have used the fact that when in \tilde{Z}_J there is a point with label 1, then the leave-one-out error is at most $\frac{1}{(n+1)^2}$ and $\beta = \mathbb{P}(\text{no point with label 1 in } \tilde{Z}_J) = \frac{\binom{2n-m}{n+1}}{\binom{2n}{n+1}}$.

The binary entropy function is concave and increasing in the interval $[0, 1/2]$. The

concavity of the binary entropy function implies that $H_b(v_2) \leq H_b(v_1)'(v_2 - v_1) + H_b(v_1)$ for all v_1 and v_2 in $(0, 1)$. Then,

$$\begin{aligned} H_b(\kappa(\tilde{Z})) &\leq H_b\left(\alpha\beta\frac{1}{n+1} + (1-\beta)\frac{1}{(n+1)^2}\right) \\ &\leq H_b\left(\frac{\alpha\beta}{n+1}\right) + \frac{1-\beta}{(n+1)^2} \log\left(\frac{n+1-\alpha\beta}{\alpha\beta}\right). \end{aligned} \quad (\text{C.38})$$

Then, by considering the summation only over \tilde{Z}_J without any point with label 1, we obtain

$$\mathbb{E}_J \sim \text{Unif}([2n]_{n+1}) \frac{1}{n+1} \sum_{j \in J} H_b(P_e(\tilde{Z}_{J-\{j\}}, \tilde{Z}_j)) \geq \alpha\beta H_b\left(\frac{1}{n+1}\right). \quad (\text{C.39})$$

Then, we can substitute Eq. (C.38) and Eq. (C.39) into Eq. (C.29) to obtain $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) = O(1)$ following the same line of reasoning as in Eq. (C.32)–Eq. (C.36).

C.8.2 Proof of Theorem C.8.1

We begin with introducing some notations. Let Γ_{2n} be the set of all bijective mappings from $[2n]$ to $\{0, 1\} \times [n]$. Let $\sigma \in \Gamma_{2n}$ and $x = (x_1, \dots, x_{2n}) \in \mathcal{X}^{2n}$ be a vector of length $2n$. Then, x^σ denotes a matrix of size $2 \times n$ where $x_{i,j}^\sigma = x_{\sigma^{-1}(i,j)}$ for $i \in \{0, 1\}$ and $j \in [n]$. Also, for every $m, k \in \mathbb{N}$ and $1 \leq k \leq m$, let $[m]_k$ denote the set of all subsets of size k of $[m]$.

For every $n \in \mathbb{N}$ let $\pi \sim \text{Unif}(\Gamma_{2n})$, $\tilde{Z} = (Z_1, \dots, Z_{2n}) \sim \mathcal{D}^{\otimes(2n)}$, and $U = (U_1, \dots, U_n) \sim (\text{Ber}(\{0, 1\}))^{\otimes n}$ where π , \tilde{Z} and U are mutually independent. Let $S_n = (\tilde{Z}_{U_j, j}^\pi)_{j=1}^n$ and $\mathcal{A}_n(S_n)$ be a learning algorithm. Let $L \in \{0, 1\}^{2 \times n}$ be a matrix with entries $L_{i,j} = \ell(\mathcal{A}_n(S_n), \tilde{Z}_{i,j}^\pi)$ for $i \in \{0, 1\}$ and $j \in [n]$. Then, using these random variables, we can define $I(L; U | \tilde{Z}, \pi)$.

Lemma C.8.2. $\text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n)) = I(L; U | \tilde{Z}, \pi)$

Proof. \tilde{Z}^π is a $\sigma(\tilde{Z}, \pi)$ -measurable random variable therefore we have $I(L; U | \tilde{Z}, \pi) = I(L; U | \tilde{Z}^\pi, \tilde{Z}, \pi)$. Then, conditioned on \tilde{Z}^π , L and U are independent from \tilde{Z} and π . Finally, note that the samples in \tilde{Z} are i.i.d., therefore we have $I(L; U | \tilde{Z}^\pi) = \text{eCMI}_{\mathcal{D}}(\ell(\mathcal{A}_n))$. \square

Lemma C.8.3. *Let π , U , and \tilde{Z} be as defined in the beginning of this section. Let $n \in \mathbb{N}$ and $f : \mathcal{Z}^n \times \mathcal{Z} \rightarrow \mathbb{R}$ be a real-valued function where f is permutation-invariant*

with respect to its first input. Then

$$\mathbb{E}^{\tilde{Z}} [f((\tilde{Z}_{U_1,1}^\pi, \dots, \tilde{Z}_{U_n,n}^\pi); \tilde{Z}_{U_1,1}^\pi)] = \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} f(\tilde{Z}_{J-\{j\}}; \tilde{Z}_j), \quad (\text{C.40})$$

where $\tilde{Z}_J = (\tilde{Z}_{J_1}, \dots, \tilde{Z}_{J_{n+1}})$ and $\bar{U}_i = 1 - U_i$.

Proof. write

$$\begin{aligned} \mathbb{E}^{\tilde{Z}} [f(\{\tilde{Z}_{U_1,1}^\pi, \dots, \tilde{Z}_{U_n,n}^\pi\}; \tilde{Z}_{U_1,1}^\pi)] &= \\ \sum_{(i_1, \dots, i_{n+1}) \in [2n]_{n+1}} f((\tilde{Z}_{i_1}, \dots, \tilde{Z}_{i_n}); \tilde{Z}_{i_{n+1}}) &\mathbb{P}^{\tilde{Z}}[(\{\tilde{Z}_{U_1,1}^\pi, \dots, \tilde{Z}_{U_n,n}^\pi\}, \tilde{Z}_{U_1,1}^\pi) = (\{\tilde{Z}_{i_1}, \dots, \tilde{Z}_{i_n}\}, \tilde{Z}_{i_{n+1}})]. \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{P}^{\tilde{Z}}[(\{\tilde{Z}_{U_1,1}^\pi, \dots, \tilde{Z}_{U_n,n}^\pi\}, \tilde{Z}_{U_1,1}^\pi) = (\{\tilde{Z}_{i_1}, \dots, \tilde{Z}_{i_n}\}, \tilde{Z}_{i_{n+1}})] &= \frac{n!(n-1)!}{(2n)!} \\ &= \frac{1}{\binom{2n}{n+1}(n+1)} \end{aligned}$$

where the last line follows since U and π are independent. It is easy to verify that is exactly equal to the RHS of Eq. (C.40). \square

Let $L = (L_1, \dots, L_n)$ where $L_i \in \{0, 1\}^2$ denotes the vector at Column i of the loss vector L . By the definition and the chain rule for mutual information we have

$$I(L; U | \tilde{Z}, \pi) = \sum_{i=1}^n I(L_i; U | L_1, \dots, L_{i-1}, \tilde{Z}, \pi) \quad (\text{C.41})$$

$$= \sum_{i=1}^n \text{H}(L_i | L_1, \dots, L_{i-1}, \tilde{Z}, \pi) - \text{H}(L_i | L_1, \dots, L_{i-1}, \tilde{Z}, \pi, U) \quad (\text{C.42})$$

$$\leq \sum_{i=1}^n \text{H}(L_i | \tilde{Z}, \pi) - \text{H}(L_i | \tilde{Z}, \pi, U) \quad (\text{C.43})$$

$$= n(\text{H}(L_1 | \tilde{Z}, \pi) - \text{H}(L_1 | \tilde{Z}, \pi, U)). \quad (\text{C.44})$$

Here, Eq. (C.43) due to the Markov chain $L_i - (U, \tilde{Z}, \pi) - L_{-i}$ and removing conditions increases the entropy. Then, the last step follows since the algorithm is permutation-invariant. Note that if \mathcal{A} is a deterministic algorithm the second term on the RHS of Eq. (C.44) is zero.

Next we provide an upper bound for $\text{H}(L_1 | \tilde{Z}, \pi)$. Note that L_1 can take values in $\{[1, 0]^\top, [0, 0]^\top, [0, 1]^\top\}$ as it is assumed that \mathcal{A}_n is consistent and, by construction,

at each column of \tilde{Z}^π one of the points is selected for the training set. Due to the symmetry imposed by π and U we have with probability one

$$\mathbb{P}^{\tilde{Z}}[L_1 = [1, 0]^\top] = \mathbb{P}^{\tilde{Z}}[L_1 = [0, 1]^\top]. \quad (\text{C.45})$$

Then, note that there exists a function κ , depending only on \mathcal{A}_n , such that

$$\kappa(\tilde{Z}) = \mathbb{P}^{\tilde{Z}}[L_{\tilde{U}_{1,1}} = 1]. \quad (\text{C.46})$$

We claim that the common value in Eq. (C.45) is given by $\frac{\kappa(\tilde{Z})}{2}$. This claim can be easily proved by taking the expectation with respect to U_1 in Eq. (C.45) and Eq. (C.46).

Then, we will show that using Lemma C.8.3, $\kappa(\tilde{Z})$ is given by the average leave-one-out error of \mathcal{A} over \tilde{Z} . Given a training set $S = ((x_1, y_1), \dots, (x_n, y_n)) \in \mathcal{Z}^n$ and a test point $z = (x, y) \in \mathcal{Z}$, let $P_e(S; z)$ denote the probability that \mathcal{A}_n makes a mistake in predicting the label x . Using this notation, we have

$$\kappa(\tilde{Z}) = \mathbb{E}^{\tilde{Z}}[P_e((\tilde{Z}_{\tilde{U}_{1,1}}^\pi, \dots, \tilde{Z}_{\tilde{U}_{n,n}}^\pi); \tilde{Z}_{\tilde{U}_{1,1}}^\pi)] \quad (\text{C.47})$$

$$= \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} P_e(\tilde{Z}_{J-\{j\}}; \tilde{Z}_j), \quad (\text{C.48})$$

where in the last step we have used Lemma C.8.3. Then, by the fact that conditioning reduces the entropy we have $\mathbb{H}(L_1 | \pi, \tilde{Z}) \leq \mathbb{H}(L_1 | \tilde{Z})$. As shown above,

$$\begin{aligned} \mathbb{P}^{\tilde{Z}}[L_1 = [1, 0]^\top] &= \mathbb{P}^{\tilde{Z}}[L_1 = [0, 1]^\top] \\ &= \frac{1}{2} \kappa(\tilde{Z}) \\ &= \frac{1}{2} \mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} P_e(\tilde{Z}_{J-\{j\}}; \tilde{Z}_j). \end{aligned} \quad (\text{C.49})$$

Then, by the fact that the L_1 can take values in $\{[1, 0]^\top, [0, 0]^\top, [0, 1]^\top\}$ and Eq. (C.49) we obtain

$$\begin{aligned} \mathbb{H}(L_1 | \tilde{Z}) &= \mathbb{E}[-(1 - \kappa(\tilde{Z})) \log(1 - \kappa(\tilde{Z})) - \kappa(\tilde{Z})/2 \log(\kappa(\tilde{Z})/2) - \kappa(\tilde{Z})/2 \log(\kappa(\tilde{Z})/2)] \\ &= \mathbb{E}[\mathbb{H}_b(\kappa(\tilde{Z})) + \kappa(\tilde{Z}) \log(2)], \end{aligned} \quad (\text{C.50})$$

where $\mathbb{H}_b(\cdot)$ denotes the binary entropy function.

Finally, we find a closed form expression for $\mathbb{H}(L_1 | \tilde{Z}, \pi, U)$. Since it is assumed that

\mathcal{A}_n is consistent we have $L_{U_i,i} = 0$ a.s.. Therefore $H(L_1|\tilde{Z}, \pi, U) = H(L_{\bar{U}_{1,1}}|\tilde{Z}, \pi, U)$. Then, we can write

$$\begin{aligned}
H(L_{\bar{U}_{1,1}}|\tilde{Z}, \pi, U) &= \mathbb{E}[H^{\tilde{Z}, \pi, U}(L_{\bar{U}_{1,1}})] \\
&= \mathbb{E}[\mathbb{E}^{\tilde{Z}}[H^{\tilde{Z}, \pi, U}(L_{\bar{U}_{1,1}})]] \\
&= \mathbb{E}[\mathbb{E}^{\tilde{Z}}\mathbb{H}_b([P_e((\tilde{Z}_{\bar{U}_{1,1}}^\pi, \dots, \tilde{Z}_{\bar{U}_{n,n}}^\pi); \tilde{Z}_{\bar{U}_{1,1}}^\pi))])] \\
&= \mathbb{E}[\mathbb{E}_{J \sim \text{Unif}([2n]_{n+1})} \frac{1}{n+1} \sum_{j \in J} \mathbb{H}_b(P_e(\tilde{Z}_{J-\{j\}}; \tilde{Z}_j)))] , \quad (\text{C.51})
\end{aligned}$$

where in the last line we have used Lemma C.8.3. Finally by combining Eq. (C.50), Eq. (C.51), and Eq. (C.44) we obtain the stated result in Eq. (C.29).

C.9 Description of the One-Inclusion Graph Prediction Algorithm

In this part we provide a short description of the one-inclusion transductive learning algorithm of Haussler, Littlestone, and M. K. Warmuth [HLW94]. Let $\bar{X} = (x_1, \dots, x_n) \in \mathcal{X}^n$ and let \mathcal{H} be a concept class with VC dimension d . Let $\mathcal{H}_{|\bar{X}}$ be the equivalence class induced by \mathcal{H} on the instances given in \bar{X} . Similarly for a classifier $h \in \mathcal{H}$, we can define $h_{|\bar{X}}$ as the restriction of h to the instances in \bar{X} . For every $h \in \mathcal{H}_{|\bar{X}}$, let $v_h = (h(x_1), \dots, h(x_n)) \in \{0, 1\}^n$. Haussler, Littlestone, and M. K. Warmuth [HLW94] defined the one-inclusion graph of \bar{X} denoted by $\mathcal{G}_{\mathcal{H}}(\bar{X}) = (V, E)$ as follows. $\mathcal{G}_{\mathcal{H}}(\bar{X})$ has the vertex set $V = \{v_h : h \in \mathcal{H}_{|\bar{X}}\}$, and $(v_h, v_{h'}) \in E$ if and only if the hamming distance of v_h and $v_{h'}$ is one. For an example of $\mathcal{G}_{\mathcal{H}}(\bar{X})$ for the concept class of intervals in one dimension, see Fig. 1 of [HLW94]. Consider *probability assignment mapping* $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})} : E \times V \rightarrow [0, 1]$ such that for each edge e incident to v_h and $v_{h'}$ the following two conditions holds.

1. for all $h'' \in \mathcal{H}_{|\bar{X}}$ with $h'' \neq h$ and $h'' \neq h'$, we have $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_{h''}) = 0$.
2. $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_h) \geq 0$, $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_{h'}) \geq 0$, and $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_h) + f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_{h'}) = 1$.

For every $i \in [n]$ and $h \in \mathcal{H}_{|\bar{X}}$, let $c_{i,h} \subset \mathcal{H}_{|\bar{X}}$ be the set of all the hypotheses in $\mathcal{H}_{|\bar{X}}$ whose restriction to $\bar{X} \setminus \{x_i\}$ equals to $h_{|\bar{X} \setminus \{x_i\}}$. Let $h^* \in \mathcal{H}$, consider a realizable $S = ((x_1, y_1), \dots, (x_n, y_n))$ where $y_i = h^*(x_i)$ for $i \in [n]$. Assume $S_{n-1} = ((x_1, y_1), \dots, (x_{n-1}, y_{n-1}))$ is given to a learner and the learner aims to predict the label of x_n . It is immediate to see that the set of hypotheses consistent with S_{n-1} is

c_{n,h^*} . Clearly, $|c_{n,h^*}| \in \{1, 2\}$ and if $|c_{n,h^*}| = 1$, we know that the target hypothesis is h^* . But, what should the learner do when $|c_{n,h^*}| = 2$?

Using $\mathcal{G}_{\mathcal{H}}(\bar{X})$ we can think of the case $|c_{n,h^*}| = 2$ as there is a vertex $v_{h'}$ adjacent to v_{h^*} , and $v_{h'}$ and v_{h^*} differ in n -th position. Assume $|c_{n,h^*}| = 2$ and $e_{n,h^*} = \{h^*, h'\}$. Using the probability assignment mapping $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}$ of $\mathcal{G}_{\mathcal{H}}(\bar{X})$, the strategy proposed by Haussler, Littlestone, and M. K. Warmuth [HLW94] predicts the label x_n to be $h'(x_n)$ with probability $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_{h^*})$, and to be $h^*(x_n)$ with probability $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_{h'})$. By identifying a deep combinatorial property of $\mathcal{G}_{\mathcal{H}}(\bar{X})$, Thm. 2.3 in [HLW94] shows that there exists a probability assignment mapping $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}$ with the mentioned properties such that $\sum_{e \in E} f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_h) \leq d$ for all $h \in \mathcal{G}_{\mathcal{H}}(\bar{X})$, and finding such a mapping is computationally easy. Moreover, Haussler, Littlestone, and M. K. Warmuth [HLW94] show there exists *deterministic* probability assignment for every one-inclusion graph such that $f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_h) \in \{0, 1\}$ for all $e \in E$ and $v_h \in V$, and $\sum_{e \in E} f_{\mathcal{G}_{\mathcal{H}}(\bar{X})}(e, v_h) \leq d$ for all $h \in \mathcal{G}_{\mathcal{H}}(\bar{X})$.

D

Appendix of Chapter 6

D.1 Proof of the information-theoretic bounds of $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ in the CLB setting

Before starting the proofs, note that the proof of Theorem 6.7.1 implies Theorem 6.4.1 (c.f. Remark 6.7.2 and Remark 6.7.5).

D.1.1 Proof of Theorem 6.7.1: Individual-sample IOMI

Consider [RBTS21, Thm. 1], which controls $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ by means of the Wasserstein distance

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{L}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbb{W}(\mathbb{P}^{Z_i}[W], \mathbb{P}(W)) \right].$$

Then, consider the fact that the Wasserstein distance is dominated by the total variation, that is, that $\mathbb{W}(\mu, \nu) \leq 2\text{RTV}(\mu, \nu)$ when the space where the distributions μ and ν are defined has diameter R with respect to the specified metric [Vil09, Thm. 6.15]¹. Applying Pinsker's [PW19, Thm. 6.5] inequality to the total variation and Jensen's inequality afterwards, one recovers the desired bound in Theorem 6.7.1.

¹In the particular case of this work, the metric considered for the Lipschitzness of the function and the diameter of the space is the ℓ_2 norm difference, but these theorems are not restricted to that.

D.1.2 Proof of Theorem 6.7.1: Individual-sample CMI

Consider now [RBTS21, Thm. 3], which again controls $\text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ by means of the Wasserstein distance

$$\text{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{L}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbb{W}(\mathbb{P}^{U_i, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}}[W], \mathbb{P}^{\tilde{Z}_{0,i}, \tilde{Z}_{1,i}}[W]) \right].$$

As in the proof above, considering the domination of the Wasserstein distance by the total variation together with Pinsker's and Jensen's inequality recovers the desired bound in Theorem 6.7.1.

D.1.3 Proof of Theorem 6.7.4

By the Donsker-Varadhan lemma [BLM13, Prop. 4.15] we have that

$$I(F, \tilde{S}; U) \geq \mathbb{E}[g(F, \tilde{S}, U)] - \log \mathbb{E} \left[e^{g(F', \tilde{S}', U)} \right]$$

for all measurable functions g such that $g(F, \tilde{S}, U)$ and $e^{g(F', \tilde{S}', U)}$ have finite expectations [BLM13, Prop. 4.15], where (F', \tilde{S}') is an independent copy of (F, \tilde{S}) and where $I(F, \tilde{S}; U) = I(F, U | \tilde{S}) = e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n))$. For the rest of the proof, let $\vec{f} \in \mathbb{R}^{2 \times n}$ be a realization of F . Consider now

$$g(\vec{f}, \tilde{s}, u) = \frac{\lambda}{n} \sum_{i=1}^n (2u_i - 1) \left(\vec{f}_{0,i} - f(0, \tilde{z}_{0,i}) - (\vec{f}_{1,i} - f(0, \tilde{z}_{1,i})) \right)$$

for some $\lambda > 0$, and note that $\mathbb{E}[g(F, \tilde{S}, U)] = \lambda \text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$. Applying Donsker-Varadhan lemma [BLM13, Prop. 4.15] with this choice of g yields

$$e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n)) \geq \lambda \text{EGE}_{\mathcal{D}}(\mathcal{A}_n) - \log \mathbb{E} \left[e^{\frac{\lambda}{n} \sum_{i=1}^n (2U_i - 1) \left(F'_{0,i} - f(0, \tilde{Z}'_{0,i}) - (F'_{1,i} - f(0, \tilde{Z}'_{1,i})) \right)} \right].$$

Studying random variables $(2U_i - 1) \left(F'_{0,i} - f(0, \tilde{Z}'_{0,i}) - (F'_{1,i} - f(0, \tilde{Z}'_{1,i})) \right)$ reveals that they are 0 mean and bounded in $[-2LR, 2LR]$. We can thus apply Hoeffding's lemma [Wai19, Example 2.4] to bound the cumulant generating function. Optimizing for $\lambda > 0$ and rearranging completes the proof.

D.2 Proof of Theorem 6.4.2

Let $d \in \mathbb{N}$ be arbitrary. Let \mathcal{W} be a ball of radius R in \mathbb{R}^d . Consider an arbitrary $z_0 \in \mathcal{W}$ such that $\|z_0\| = R$. The input space is $\mathcal{Z} = \{z_0/R, -z_0/R\}$. Also, let the data distribution \mathcal{D} be $\mathcal{D}(z_0/R) = \mathcal{D}(-z_0/R) = 1/2$, the loss function $f : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}$ be $f(w, z) = -L \langle w, z \rangle$. It is straightforward to see that the loss function is convex and L -Lipschitz ².

Denote the training set $S = (Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$. Define a Rademacher random variable $\epsilon_i = 1$ if $Z_i = z_0/R$ and $\epsilon_i = -1$ if $Z_i = -z_0/R$. We can represent the training set as $S = (\frac{z_0}{R}\epsilon_1, \dots, \frac{z_0}{R}\epsilon_n)$. The empirical risk for $w \in \mathcal{W}$ is given by $\hat{F}_{S_n}(w) = \frac{-L}{nR} \langle w, z_0 \sum_{i \in [n]} \epsilon_i \rangle$. It is straightforward to see that the ERM for this problem is

$$\arg \min_{w \in \mathcal{W}} \hat{F}_{S_n}(w) = \mathcal{A}_n(S) = \begin{cases} z_0 & \text{if } \text{sign}(\sum_{i=1}^n \epsilon_i) = 1 \\ -z_0 & \text{if } \text{sign}(\sum_{i=1}^n \epsilon_i) = -1 \end{cases},$$

where for $x \in \mathbb{R}$, $\text{sign}(x) = 1$ if $x \geq 0$ and $\text{sign}(x) = -1$ if $x < 0$.

First, we provide a lower bound on the expected generalization error. The expected empirical risk of \mathcal{A}_n is given by

$$\begin{aligned} \mathbb{E} \left[\min_{w \in \mathcal{W}} \hat{F}_{S_n}(w) \right] &= \mathbb{E} \left[\min_{w \in \mathcal{W}} -\frac{L}{Rn} \left\langle w, z_0 \sum_{i=1}^n \epsilon_i \right\rangle \right] \\ &= -\frac{L}{Rn} \mathbb{E} \left[\max_{w \in \{z_0, -z_0\}} \left\langle w, z_0 \sum_{i=1}^n \epsilon_i \right\rangle \right] \\ &= -\frac{L}{Rn} \mathbb{E} \left[\left| \left\langle z_0, z_0 \sum_{i=1}^n \epsilon_i \right\rangle \right| \right] \\ &= -\frac{LR}{n} \mathbb{E} \left[\left| \sum_{i=1}^n \epsilon_i \right| \right], \end{aligned}$$

where we have used $\forall a, b \in \mathbb{R}$, $\max(a, b) = \frac{a+b}{2} + \frac{|a-b|}{2}$. Observe that $F_{\mathcal{D}}(w) = 0$ for

²The construction for this section is inspired by the lower bounds for online convex optimization in [Ora19, Sec.5.1].

all $w \in \mathcal{W}$. Therefore, the expected generalization error is lower bounded by

$$\begin{aligned} \text{EGE}_{\mathcal{D}}(\mathcal{A}_n) &= -\mathbb{E} \left[\min_{w \in \mathcal{W}} \hat{F}_{S_n}(w) \right] \\ &= \frac{LR}{n} \mathbb{E} \left[\left| \sum_{i=1}^n \epsilon_i \right| \right] \\ &\geq \frac{LR}{\sqrt{2n}}, \end{aligned}$$

where the last line follows from Khintchine–Kahane inequality [MRT18, Thm. D.9].

Next, we analyze the upper bounds based on Theorem 6.4.1. Observe that the following Markov chain holds:

$$S - \text{sign} \left(\sum_{i=1}^n \epsilon_i \right) - \mathcal{A}_n(S).$$

By the data processing inequality we have

$$\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n) = I(\mathcal{A}_n(S); S) \leq I(\mathcal{A}_n(S); \text{sign} \left(\sum_{i=1}^n \epsilon_i \right)).$$

We can upper bound the mutual information as

$$I(\mathcal{A}_n(S); \text{sign} \left(\sum_{i=1}^n \epsilon_i \right)) \leq H(\text{sign} \left(\sum_{i=1}^n \epsilon_i \right)) \leq 1,$$

since $\text{sign} \left(\sum_{i=1}^n \epsilon_i \right)$ can take only two values. Therefore, we obtain $\text{IOMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 1$. As $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \text{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ for any learning problem [HNKRD20, Thm. 2.1], we have $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 1$. Finally, the result follows by plugging the bounds on IOMI and CMI into Theorem 6.4.1.

D.3 Proof of Theorem 6.5.1

The outline of the proof is as follows. First, in Appendix D.3.1, we describe our construction. Then, we analyze the dynamics of GD on the problem in Appendix D.3.2. Using the properties of the final iterate of GD, proved in Appendix D.3.2, we proceed by showing in Appendix D.3.3 that if the noise variance is greater than a threshold, then the residual term does not converge to zero as the number of samples grows. For the case that the noise variance is smaller than the threshold, we prove the failure of

IOMI and CMI in Appendix D.3.4 and Appendix D.3.5, respectively.

D.3.1 Construction

We begin the proof by describing a learning scenario that witnesses the lower bound (we drop the n argument from the parameters to reduce notational clutter). Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{0, 1\}^d$. Let the data distribution on input be $(\text{Ber}(1/2))^{\otimes d}$, i.e., each coordinate is drawn independently and uniformly at random from $\text{Ber}(1/2)$. In this section, we treat the training set $S \in \{0, 1\}^{n \times d}$ as a matrix. Note that each element of S is drawn i.i.d. from $\text{Ber}(1/2)$. For $i \in [d]$, we say the i -th coordinate is a *bad coordinate* iff for all $j \in [n]$, $Z_j(i) = 0$. In words, if i -th coordinate is a bad coordinate then all the entries in the i -th column of S is zero. Also, the convex domain space \mathcal{W} is the Euclidean ball of radius one in \mathbb{R}^d . Note for $x \in \mathbb{R}^d$, $\Pi_{\mathcal{W}}(x) = x / \max\{\|x\|, 1\}$.

We consider the convex function proposed in Amir, Koren, and Livni [AKL21]. Let $0 < \lambda \leq \mathcal{O}(1/(n\sqrt{d}))$ be a positive constant which is determined later. Then we consider the following loss function $f : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}$

$$f(w, z) = \sum_{i=1}^d z(i)w(i)^2 + \lambda \langle w, z \rangle + \max \left\{ \max_{i \in [d]} \{w(i)\}, 0 \right\}. \quad (\text{D.1})$$

It is straightforward to show that the first two terms in Eq. (D.1) is convex. Also, $\max\{\max_{i \in [d]} \{w(i)\}, 0\}$ is a convex function because it is maximum of convex (linear) functions [BV04, Sec.3.2.3]. Therefore, f is convex as it is sum of convex functions. Then, we show that each term in Eq. (D.1) is Lipschitz. The first term, $\sum_{i=1}^d z(i)w(i)^2 \leq \|w\|^2$ is 2-Lipschitz by the boundedness of \mathcal{W} . The second term is $\lambda\sqrt{d}$ -Lipschitz because $\|\nabla(\lambda \langle w, z \rangle)\| = \lambda\|z\| \leq \lambda\sqrt{d}$. We use Lemma D.6.5 to show that the last term in Eq. (D.1) is 1-Lipschitz. Therefore, $f(w, z)$ is $(3 + \lambda\sqrt{d})$ -Lipschitz. Note that $\lambda \in \mathcal{O}(1/(n\sqrt{d}))$, so the function in Eq. (D.1) is 4-Lipschitz for sufficiently large n .

D.3.2 Dynamics of GD

First of all, we want to note that the statements in this proof about random variables hold almost surely. We will skip such declarations for the remainder of the proof to aid readability. In this part, we aim to find the properties of the final iterates of the GD algorithm. Let $d = 0.75T2^n$. Let $\mathbf{B} \in \{0, 1\}^d$ denote a vector such that $\mathbf{B}(i) = 1$ if and only if i is a bad coordinate. Let $\|\mathbf{B}\|_0$ denote the number of bad coordinates.

Next, we provide a probabilistic estimate on $\|\mathbf{B}\|_0$. $\|\mathbf{B}\|_0 = \sum_{i=1}^d \mathbf{B}(i)$ follows the binomial distribution with the number of trial d and the success probability of 2^{-n} . The reason is the probability that all the points in a column is zero is given by 2^{-n} . By the standard multiplicative Chernoff bound [MU05, Cor.4.6] we have

$$\mathbb{P}(T/2 \leq \|\mathbf{B}\|_0 \leq T) \geq 1 - 2 \exp(-T/36). \quad (\text{D.2})$$

Therefore, with probability at least $1 - 2 \exp(-T/36)$, the number of bad coordinates is between $T/2$ and T .

Next step concerns understanding the dynamics of GD. The empirical risk for any $w \in \mathcal{W}$ is given by

$$\hat{\mathbb{F}}_{S_n}(w) = \sum_{i=1}^d \hat{\mu}(i) w(i)^2 + \lambda \langle \hat{\mu}, w \rangle + \max \left\{ \max_{i \in [d]} \{w(i)\}, 0 \right\} \quad (\text{D.3})$$

where for $i \in [d]$, $\hat{\mu}(i) = \frac{1}{n} \sum_{j=1}^n z_j(i) \in [0, 1]$ is the empirical mean of the points in i -th column of S .

Lemma D.3.1. *Under the event $\{T/2 \leq \|\mathbf{B}\|_0 \leq T\}$, let $\mathcal{B} = \{i_1, \dots, i_{\|\mathbf{B}\|_0}\} \subseteq [d]$ contain the ordered set of bad coordinates. Consider the GD process $W_{t+1} = W_t - \Pi_{\mathcal{W}}(W_t - \eta \partial(\hat{\mathbb{F}}_{S_n}(W_t)))$ starting at $W_0 = 0$ where η is the step size. For every $i \in [d]$ and $t \in [T]$*

$$W_t(i) = \begin{cases} \frac{\lambda}{2}(-1 + (1 - 2\eta\hat{\mu}(i))^t) & i \in [d] \setminus \mathcal{B} \\ -\eta & i \in \{i_1, \dots, i_{\min\{\|\mathbf{B}\|_0, t-1\}}\} \\ 0 & i \in \{i_{\min\{\|\mathbf{B}\|_0, t-1\}+1}, \dots, i_{\|\mathbf{B}\|_0}\} \end{cases}.$$

In particular,

$$W_T(i) = \begin{cases} \frac{\lambda}{2}(-1 + (1 - 2\eta\hat{\mu}(i))^T) & i \in [d] \setminus \mathcal{B} \\ -\eta & i \in \mathcal{B} \end{cases},$$

and for all $i \in [d] \setminus \mathcal{B}$, $-\eta\lambda T \leq W_T(i) < 0$.

Proof. First, we describe the first-order oracle proposed in Amir, Koren, and Livni [AKL21] and Bassily, Feldman, Guzmán, and Talwar [BFGT20]. Note that the first two terms in Eq. (D.1) are differentiable. For the third term, i.e., $f_3(w) = \max\{\max_{i \in [d]} \{w(i)\}, 0\}$, which is not differentiable we consider the following first-order oracle. Let $\mathcal{I}(w) = \{j \in [d] | j \in \{\arg \max_{i \in [d]} w(i)\} \cap \{i | w(i) \geq 0\}\}$. Then, we

claim that

$$\partial f_3(w) = \begin{cases} 0 & w = 0 \text{ or } \mathcal{I} = \emptyset \\ \mathbf{e}(\min\{\mathcal{I}(w)\}) & w \neq 0 \text{ and } \mathcal{I} \neq \emptyset. \end{cases} \quad (\text{D.4})$$

where for $i \in [d]$, $\mathbf{e}(i) = (\underbrace{0, \dots, 0}_{i-1 \text{ times}}, 1, \underbrace{0, \dots, 0}_{d-i \text{ times}})$ (i -th coordinate vector).

To prove that Eq. (D.4) is a member of subgradient at w , we need to prove for all $w, v \in \mathbb{R}^d$, we have $f_3(v) \geq f_3(w) + \langle \partial f_3(w), v - w \rangle$.

Consider the case $w = 0$, then, since $\partial f_3(w) = 0$, trivially we have $f_3(v) \geq f_3(w) = 0$. Next, consider the case that $w \neq 0$ but $\mathcal{I}(w) = \emptyset$. This case holds if and only if for all $i \in [d]$, $w(i) < 0$. Therefore, $\partial f_3(w) = 0$ and the first-order convexity condition trivially holds. Finally, consider the case that $w \neq 0$ and $\mathcal{I}(w) \neq \emptyset$. Let $\hat{i} = \min\{\mathcal{I}(w)\}$, then

$$\begin{aligned} f_3(w) + \langle \mathbf{e}(\hat{i}), v - w \rangle &= w(\hat{i}) + \langle \mathbf{e}(\hat{i}), v - w \rangle \\ &= w(\hat{i}) + v(\hat{i}) - w(\hat{i}) \\ &= v(\hat{i}) \\ &\leq \max \left\{ \max_{i \in [d]} \{v(i)\}, 0 \right\}, \end{aligned}$$

as was to be shown.

Then, we provide analysis of the dynamics of GD using the first-order oracle described above. We only describe the dynamics under the event $\{T/2 \leq \|\mathbf{B}\|_0 \leq T\}$. Let the (ordered) set of bad coordinates denoted by $\mathcal{B} = \{i_1, \dots, i_{\|\mathbf{B}\|_0}\}$. The main observation here is that we can re-write the Eq. (D.3) as follows

$$\hat{F}_{S_n}(w) = \sum_{i \in [d] \setminus \mathcal{B}} w(i)^2 \hat{\mu}(i) + \lambda \sum_{i \in [d] \setminus \mathcal{B}} w(i) \hat{\mu}(i) + \max \left\{ \max_{i \in [d]} \{w(i)\}, 0 \right\}. \quad (\text{D.5})$$

This equation shows that the gradient comes from the first two terms does not change the bad coordinates of w . As we will show that f_3 does not provide gradient for bad coordinates, the dynamic of each good coordinate of w is independent of other

coordinates. Formally, we prove by induction that $W_1 = -\eta\lambda\hat{\mu}$, and for $t \geq 2$,

$$W_t(i) = \begin{cases} \frac{\lambda}{2}(-1 + (1 - 2\eta\hat{\mu}(i))^t) & i \in [d] \setminus \mathcal{B} \\ -\eta & i \in \{i_1, \dots, i_{\min\{\|\mathbf{B}\|_0, t-1\}}\} \\ 0 & i \in \{i_{\min\{\|\mathbf{B}\|_0, t-1\}+1}, \dots, i_{\|\mathbf{B}\|_0}\} \end{cases}.$$

For the base case, by the GD algorithm's update rule we have $W_1 = \Pi_{\mathcal{W}}(W_0 - \eta g_0) = \Pi_{\mathcal{W}}(-\eta g_0)$. Note that $g_0 = \lambda\hat{\mu}$. Since $\lambda \in \mathcal{O}(1/(n\sqrt{d}))$, $-\eta g_0 \in \mathcal{W}$.

For the inductive step, assume that for some $k \in [T-1]$, the claim holds. We have $W_{k+1} = \Pi_{\mathcal{W}}(W_k - \eta g_k)$. First, for $i \in [d] \setminus \mathcal{B}$, $g_k(i) = 2W_k(i)\hat{\mu}(i) + \lambda\hat{\mu}(i)$. Note that the gradient from the third term is zero for good coordinates as $W_k(i) < 0$ for $i \in [d] \setminus \mathcal{B}$. By a simple calculation, one can show that

$$W_k - \eta g_k = \frac{\lambda}{2}(-1 + (1 - 2\eta\hat{\mu}(i))^{k+1}).$$

Also, as $\lambda \in \mathcal{O}(1/(n\sqrt{d}))$, $W_k - \eta g_k \in \mathcal{W}$. Then, for the bad coordinates, consider two cases $\min\{\|\mathbf{B}\|_0, k-1\} = k-1$ and $\min\{\|\mathbf{B}\|_0, k-1\} = \|\mathbf{B}\|_0$. Consider the first case, i.e., $\min\{\|\mathbf{B}\|_0, k-1\} = k-1$. Consider $i_k \in \mathcal{B}$. From Eq. (D.5), the first two terms do not provide gradient for bad coordinates. Then, we claim that $\partial f_3(W_k) = \mathbf{e}(i_k)$. The reason is that for all $i \in \{i_1, \dots, i_{k-1}\} \cup [d] \setminus \mathcal{B}$, $W_k(i) < 0$ and $i \in \{i_k, \dots, \|\mathbf{B}\|_0\}$, $W_k(i) = 0$. Therefore, the claim follows from Eq. (D.4). Therefore, for the first case, for all $i \in \{i_1, \dots, i_k\}$, $W_{k+1}(i) = -\eta$, and for all $i \in \{i_{k+1}, \dots, \|\mathbf{B}\|_0\}$, $W_{k+1}(i) = 0$.

Consider the second case, $\min\{\|\mathbf{B}\|_0, k-1\} = \|\mathbf{B}\|_0$. In this case, all coordinates of W_k are less than zero. Therefore, the gradient from f_3 is zero, and the bad coordinates remain unchanged. \square

Next, we provide a result regarding $\|W_T\|$.

Lemma D.3.2. *Under the event $\{T/2 \leq \|\mathbf{B}\|_0 \leq T\}$, we have $\frac{1}{2\sqrt{n}} \leq \|W_T\| \leq \frac{1}{\sqrt{n}}$.*

Proof. Under the event $\{T/2 \leq \|\mathbf{B}\|_0 \leq T\}$, Lemma D.3.1 shows that

$$\|W_T\| = \left(\|\mathbf{B}\|_0 \eta^2 + \sum_{i \in [d] \setminus \mathcal{B}} W_T(i)^2 \right)^{\frac{1}{2}}.$$

Since for good coordinates, $|W_T(i)| \leq \lambda\eta T$, we have the following upper bound $\|W_T\| \leq \sqrt{T\eta^2 + d(\lambda\eta T)^2}$. For a lower bound consider $\|W_T\| \geq \sqrt{T\eta^2/2}$. Setting the parameters, we obtain $\|W_T\| \leq \frac{1}{\sqrt{n}}$ and $\|W_T\| \geq \frac{1}{2\sqrt{n}}$. \square

D.3.3 Noise with Large Variance Fails

Consider the case that the variance of ξ along each dimension is σ^2 and $\sigma \geq \frac{\beta^*}{\sqrt{d}}$ where $\beta^* = 0.1$. In particular, $\frac{\beta^*}{\sqrt{d}}$ is the threshold for the variance. First of all note that for all $w \in \mathcal{W}$

$$F_{\mathcal{D}}(w) = \frac{1}{2}\|w\|^2 + \frac{\lambda}{2} \sum_{i=1}^d w(i) + \max \left\{ \max_{i \in [d]} \{w(i)\}, 0 \right\}.$$

Therefore,

$$|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| = \left| \frac{1}{2}(\|\tilde{W}_T\|^2 - \|W_T\|^2) + \frac{\lambda}{2} \sum_{i=1}^d (\tilde{W}_T(i) - W_T(i)) + \Xi_T \right|, \quad (\text{D.6})$$

where $\Xi_T = \max\{\max_{i \in [d]} \{\tilde{W}_T(i)\}, 0\} - \max\{\max_{i \in [d]} \{W_T(i)\}, 0\}$. Under the event $\{T/2 \leq \|\mathbf{B}\|_0 \leq T\}$, Lemma D.3.1 shows that $\max\{\max_{i \in [d]} \{W_T(i)\}, 0\} = 0$ since $W_T(i) < 0$ for all $i \in [d]$. Therefore, $\Xi_T = \max\{\max_{i \in [d]} \{\tilde{W}_T(i)\}, 0\} \geq 0$.

Because the Gaussian distribution is invariant under the rotation, we can assume that $W_T = (\|W_T\|, \underbrace{0, \dots, 0}_{d-1 \text{ times}})$ without loss of generality. Therefore, Eq. (D.6) is given by

$$|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| = \left| \frac{1}{2}(\|\tilde{W}_T\|^2 - \|W_T\|^2) + \frac{\lambda}{2}(\tilde{W}_T(1) - \|W_T\|) + \sum_{i=2}^d \tilde{W}_T(i) + \Xi_T \right|. \quad (\text{D.7})$$

Let $V_T = W_T + \xi$. Let us represent $\xi = r\theta$ where $r = \|\xi\|$ and $\theta = \xi/\|\xi\|$. By a simple calculation, one can obtain that

$$\|V_T\|^2 = \|W_T\|^2 + r^2 + 2\|W_T\|r\theta(1). \quad (\text{D.8})$$

Define $\mathbf{r}_{\max} = 1 - \|W_T\|$. Note $0 \leq \mathbf{r}_{\max} \leq 1$ since $W_T \in \mathcal{W}$. By the tower rule for the expectation,

$$\begin{aligned} \mathbb{E} \left[|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| \right] &\geq \mathbb{E} \left[|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T] \mathbf{1}[r \leq \mathbf{r}_{\max}] \right] \\ &+ \mathbb{E} \left[|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T] \mathbf{1}[r > \mathbf{r}_{\max}] \mathbf{1}[\|V_T\| \leq 1] \right] \\ &+ \mathbb{E} \left[|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T] \mathbf{1}[r > \mathbf{r}_{\max}] \mathbf{1}[\|V_T\| > 1] \right] \end{aligned} \quad (\text{D.9})$$

Under the event $\{T/2 \leq \|\mathbf{B}\|_0 \leq T\}$, we divide the sample space into three regions: **Region 1**: $\{r \leq r_{\max}\}$, **Region 2**: $\{r > r_{\max}\} \cap \{\|V_T\| < 1\}$, and **Region 3**: $\{r > r_{\max}\} \cap \{\|V_T\| \geq 1\}$. In what follows, we lower bound Eq. (D.9) for each region separately.

[R_1] **Region 1** $\{r \leq r_{\max}\}$:

By the tower rule for the expectation,

$$\begin{aligned} & \mathbb{E} \left[|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T] \mathbf{1}[r \leq r_{\max}] \right] \\ &= \mathbb{E} \left[\mathbb{E}^{W_T, r} \left[|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| \right] \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T] \mathbf{1}[r \leq r_{\max}] \right] \end{aligned}$$

Under the event $\{r \leq r_{\max}\}$, it is straightforward to see that $\|V_T\| \leq 1$. Therefore, $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T) = V_T$, and Eq. (D.7) is given by

$$|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| = \left| \frac{1}{2}r^2 + \|W_T\|r\theta(1) + \frac{\lambda r}{2} \sum_{i=1}^d \theta(i) + \Xi_T \right|.$$

By the construction of the surrogate algorithm and Lemma D.6.1, we know that θ is independent of r and W_T . Then, we invoke the reverse triangle inequality, i.e., $|a - b| \geq |a| - |b|$ for $a, b \in \mathbb{R}$. Using $\theta \stackrel{d}{=} -\theta$, we have

$$\begin{aligned} & \mathbb{E}^{r, W_T} \left[\left| \frac{1}{2}r^2 + r\|W_T\|\theta(1) + \frac{\lambda r}{2} \sum_{i=1}^d \theta(i) + \Xi_T \right| \right] \\ & \geq \underbrace{\mathbb{E}^{W_T, r} \left[\left| \frac{1}{2}r^2 + \|W_T\|r\theta(1) + \Xi_T \right| \right]}_{\textcircled{1}} - \underbrace{\mathbb{E}^{W_T, r} \left[\left| \frac{\lambda r}{2} \sum_{i=1}^d \theta(i) \right| \right]}_{\textcircled{2}}. \end{aligned} \quad (\text{D.10})$$

We will analyze $\textcircled{1}$ and $\textcircled{2}$ separately. Note that $\theta(1) \sim \text{Unif}([-1, 1])$. Thus, with probability $1/2$, $\theta(1) \in [0, 1]$. Therefore,

$$\textcircled{1} \geq \mathbb{E}^{W_T, r} \left[\left| \frac{1}{2}r^2 + \|W_T\|r\theta(1) + \Xi_T \right| \mathbf{1}[\theta(1) \in [0, 1]] \right] \geq \left| \frac{r^2}{4} + \frac{\Xi_T}{2} \right| \geq \frac{r^2}{4}, \quad (\text{D.11})$$

where the last inequality follows from $\Xi_T \geq 0$. By the Cauchy-Schwartz, we have

$\|\theta\|_1 \leq \sqrt{d}$ since $\|\theta\|_2 = 1$. Therefore,

$$\begin{aligned}
\textcircled{2} &\leq \frac{\lambda r}{2} \|\theta\|_1 \\
&\leq \frac{\lambda r}{2} \sqrt{d} \\
&\leq \frac{\lambda r_{\max}}{2} \sqrt{d} \\
&\leq \frac{\lambda}{2} \sqrt{d},
\end{aligned} \tag{D.12}$$

where the third inequality follows since $r \leq r_{\max}$ and the the last step follows from r_{\max} being less than one. By Eq. (D.10), Eq. (D.11), and Eq. (D.12), we finish lower bounding the inner expectation,

$$\mathbb{E}^{W_T, r} \left[|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| \right] \geq \frac{r^2}{4} - \frac{\lambda \sqrt{d}}{2} \geq \frac{r^2}{4} - \frac{1}{2n}. \tag{D.13}$$

Here, the last inequality follows from setting $\lambda \leq \frac{1}{n\sqrt{d}}$.

[R₂] **Region 2:** $\{r > r_{\max}\}$ and $\{\|V_T\| < 1\}$:

Since $\|V_T\| < 1$ and $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T)$, we have $\tilde{W}_T = V_T$. Using Eq. (D.8), we can write

$$|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| \geq \left| \frac{1}{2}(r^2 + 2\|W_T\|r\theta(1)) + \frac{\lambda r}{2} \sum_{i \in [d]} \theta(i) + \Xi_T \right|.$$

Then, using the reverse triangle inequality, i.e., $|a - b| \geq |a| - |b|$ for $a, b \in \mathbb{R}$, and the facts that $|\theta(1)| \leq 1$ and $\Xi_T \geq 0$, we have

$$\begin{aligned}
|F_{\mathcal{D}}(\tilde{W}_T) - F_{\mathcal{D}}(W_T)| &\geq \left| \frac{1}{2}r^2 + \Xi_T \right| - \left| r\|W_T\|\theta(1) + \frac{\lambda r}{2} \sum_{i \in [d]} \theta(i) \right| \\
&\geq \left| \frac{1}{2}r^2 + \Xi_T \right| - r\|W_T\||\theta(1)| - \frac{\lambda r}{2} \left| \sum_{i \in [d]} \theta(i) \right| \\
&\geq \frac{1}{2}r^2 + \Xi_T - r\|W_T\| - \frac{\lambda r}{2} \left| \sum_{i \in [d]} \theta(i) \right| \\
&\geq \frac{1}{2}r^2 - r\|W_T\| - \frac{\lambda r}{2} \left| \sum_{i \in [d]} \theta(i) \right|.
\end{aligned}$$

By the Cauchy-Schwartz, we have $\|\theta\|_1 \leq \sqrt{d}$ since $\|\theta\|_2 = 1$. Therefore,

$$\begin{aligned} |\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| &\geq \frac{1}{2}r^2 - r\|W_T\| - \frac{\lambda r}{2}\sqrt{d} \\ &\geq \frac{1}{2}r^2 - r\|W_T\| - \frac{r}{2n}. \end{aligned} \quad (\text{D.14})$$

Here, the last line follows from $\lambda \leq \frac{1}{n\sqrt{d}}$.

Define $g : \mathbb{R} \rightarrow \mathbb{R}$ where $g(x) = x^2/2 - x(\|W_T\| + 1/(2n))$. Then, we have $\arg \min_{x \in \mathbb{R}} g(x) = \|W_T\| + 1/(2n)$. From Lemma D.3.2, we know that $\|W_T\| \leq 1/\sqrt{n}$. Notice that for $n \geq 5$, we have $\|W_T\| \leq 1/\sqrt{n} \leq 0.5(1 - 1/(2n))$ which gives us $\|W_T\| + 1/(2n) \leq 1 - \|W_T\| = r_{\max}$. Therefore, we conclude that g is increasing for $x \geq r_{\max}$. Note that the lower bound in Eq. (D.14) is $g(r)$ and using this observation we have $g(r) > g(r_{\max})$ since in this region $r > r_{\max}$. Therefore, we can further lower bound Eq. (D.14) as

$$\begin{aligned} |\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| &\geq \frac{3}{2}\|W_T\|^2 - \left(2 - \frac{1}{2n}\right)\|W_T\| + \frac{1}{2}\left(1 - \frac{1}{n}\right) \\ &\geq \frac{1}{2} - \frac{2}{\sqrt{n}}. \end{aligned} \quad (\text{D.15})$$

To prove the last step define $h : \mathbb{R} \rightarrow \mathbb{R}$ where $h(x) = \frac{3}{2}x^2 - \left(2 - \frac{1}{2n}\right)x + \frac{1}{2}\left(1 - \frac{1}{n}\right)$. It is straightforward to see that $h(x)$ is decreasing for $x \leq 1/\sqrt{n}$ when $n \geq \sqrt{5}$. Using this argument and some manipulations we can show the last step.

[R₃] **Region 3:** $\{r > r_{\max}\}$ and $\{\|V_T\| \geq 1\}$.

Since $\|V_T\| \geq 1$ and $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T)$, we have $\|\tilde{W}_T\| = 1$. Using this observation and reverse triangle inequality, i.e., $|a - b| \geq |a| - |b|$ for $a, b \in \mathbb{R}$, we can simplify Eq. (D.7) as

$$\begin{aligned} |\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| &\geq \frac{1}{2}|1 - \|W_T\|^2 + 2\Xi_T| - \frac{\lambda}{2} \left| \sum_{i \in [d]} \tilde{W}_T(i) - \|W_T\| \right| \\ &\geq \frac{1}{2}|1 - \|W_T\|^2 + 2\Xi_T| - \frac{\lambda}{2}(\|\tilde{W}_T\|_1 + \|W_T\|). \end{aligned}$$

The last line follows from using the triangle inequality twice. By Lemma D.3.2, we have $\|W_T\| \leq 1/\sqrt{n}$. Then, since $\Xi_T \geq 0$, we obtain

$$|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| \geq \frac{1}{2}\left(1 + 2\Xi_T - \frac{1}{n}\right) - \frac{\lambda}{2}\left(\|\tilde{W}_T\|_1 + \frac{1}{\sqrt{n}}\right).$$

Also, by the Cauchy-Schwartz, we have $\|\tilde{W}_T\|_1 \leq \sqrt{d}$ since $\|\tilde{W}_T\|_2 = 1$. Therefore, setting $\lambda \leq \frac{1}{n\sqrt{d}}$

$$\begin{aligned} |\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| &\geq \frac{1}{2}\left(1 - \frac{1}{n}\right) + \Xi_T - \frac{1}{2n} - \frac{1}{2n^{1.5}\sqrt{d}} \\ &\geq \frac{1}{2} - \frac{1}{n} \\ &\geq \frac{1}{2} - \frac{2}{\sqrt{n}}. \end{aligned} \quad (\text{D.16})$$

Here the last line follows from $\Xi_T \geq 0$ and some simple manipulations.

Equipped with the lower bounds for each region we can conclude this part of the proof. Combining Eq. (D.9) with Eq. (D.13), Eq. (D.15), and Eq. (D.16), we obtain

$$\begin{aligned} \mathbb{E} \left[|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| \right] &\geq \mathbb{E} \left[\left(\frac{r^2}{4} - \frac{1}{2n} \right) \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T] \mathbf{1}[r \leq r_{\max}] \right] \\ &+ \left(\frac{1}{2} - \frac{2}{\sqrt{n}} \right) \mathbb{E} \left[\mathbb{P}^S[r > r_{\max}] \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T] \right]. \end{aligned} \quad (\text{D.17})$$

Assume we choose n sufficiently large so that $\frac{(\beta^*)^2}{16} - \frac{1}{2n} \geq 0$ (Notice that such n always exists). We can further lower bound Eq. (D.17) as

$$\begin{aligned} \mathbb{E} \left[|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)| \right] &\geq \mathbb{E} \left[\left(\frac{r^2}{4} - \frac{1}{2n} \right) \mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \mathbf{1}\left[\frac{\beta^*}{2} \leq r \leq r_{\max}\right] \right] \\ &+ \mathbb{E} \left[\left(\frac{r^2}{4} - \frac{1}{2n} \right) \mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \mathbf{1}\left[r < \frac{\beta^*}{2}\right] \right] \\ &+ \left(\frac{1}{2} - \frac{2}{\sqrt{n}} \right) \mathbb{E} \left[\mathbb{P}^S[r > r_{\max}] \mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \right] \\ &\geq \left(\frac{(\beta^*)^2}{16} - \frac{1}{2n} \right) \mathbb{E} \left[\mathbb{P}^S\left[\frac{\beta^*}{2} \leq r \leq r_{\max}\right] \mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \right] \\ &- \frac{1}{2n} \mathbb{E} \left[\mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \mathbf{1}\left[r < \frac{\beta^*}{2}\right] \right] + \left(\frac{1}{2} - \frac{2}{\sqrt{n}} \right) \mathbb{E} \left[\mathbb{P}^S[r > r_{\max}] \mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \right] \\ &\geq \left(\frac{(\beta^*)^2}{16} - \frac{1}{2n} \right) \mathbb{E} \left[\mathbb{P}^S\left[\frac{\beta^*}{2} \leq r\right] \mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \right] - \frac{1}{2n} \mathbb{E} \left[\mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \mathbf{1}\left[r < \frac{\beta^*}{2}\right] \right] \\ &\geq \left(\frac{(\beta^*)^2}{16} - \frac{1}{2n} \right) \mathbb{E} \left[\mathbb{P}^S\left[\frac{\beta^*}{2} \leq r\right] \mathbf{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right] \right] - \frac{1}{2n} \mathbb{P}(r < \frac{\beta^*}{2}). \end{aligned} \quad (\text{D.18})$$

Here, we have used $\frac{1}{2} - \frac{2}{\sqrt{n}} \geq \frac{(\beta^*)^2}{16} - \frac{1}{2n}$ for $n \geq 14$ where $\beta^* = 0.1$, and $\frac{r^2}{4} - \frac{1}{2n} \geq -\frac{1}{2n}$ for $r \geq 0$.

Note that $S \perp\!\!\!\perp r$ by the construction of the surrogate algorithm. By assumption

we have $\sigma \geq \frac{\beta^*}{\sqrt{d}}$. Using the concentration bound from Corollary D.6.3 we obtain

$$\mathbb{P}\left(r \leq \frac{\beta^*}{2}\right) \leq \mathbb{P}\left(r \leq \frac{\sigma\sqrt{d}}{2}\right) \leq 2 \exp\left(-\frac{9d}{64}\right).$$

Since r and S are independent, by Eq. (D.2) we have

$$\begin{aligned} \mathbb{E}\left[\mathbb{P}^S\left[\frac{\beta^*}{2} \leq r\right] \mathbb{1}\left[\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right]\right] &= \mathbb{P}\left(\frac{\beta^*}{2} \leq r\right) \mathbb{P}\left(\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right) \\ &\geq (1 - 2 \exp(-9d/64)) \mathbb{P}\left(\frac{T}{2} \leq \|\mathbf{B}\|_0 \leq T\right) \\ &\geq (1 - 2 \exp(-9d/64))(1 - 2 \exp(-T/36)). \end{aligned} \tag{D.19}$$

Therefore, we conclude this part by combining Eq. (D.18) and Eq. (D.19) to obtain the following lower bound:

$$\begin{aligned} &\mathbb{E}\left[|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)|\right] \\ &\geq \left(\frac{(\beta^*)^2}{16} - \frac{1}{2n}\right) (1 - 2 \exp(-9d/16) - 2 \exp(-T/36)) - \frac{1}{n} \exp(-\frac{9d}{64}). \end{aligned}$$

By setting the parameters, i.e., T and d , we prove that for sufficiently large n

$$\mathbb{E}\left[|\mathbb{F}_{\mathcal{D}}(\tilde{W}_T) - \mathbb{F}_{\mathcal{D}}(W_T)|\right] \in \Omega(1),$$

which was to be shown.

D.3.4 Noise With Small Variance Fails: IOMI

In Appendix D.3.3 we showed that if the variance of ξ is greater than $(\frac{\beta^*}{\sqrt{d}})^2$, the distance between the population risk of the surrogate algorithm and the GD algorithm does not go to zero. In this part, we will show that if the variance of ξ is smaller than $\frac{\beta^*}{\sqrt{d}}$ then, the mutual information term does not vanish as $n \rightarrow \infty$.

By the definition of the mutual information we can write $I(\tilde{W}_T; S) = \mathbb{H}(S) - \mathbb{H}(S|\tilde{W}_T)$. Note that \mathbf{B} is a S -measurable random variable. Therefore, we have

$$\mathbb{H}(S|\tilde{W}_T) = \mathbb{H}(S, \mathbf{B}|\tilde{W}_T).$$

Then, by the chain rule for the discrete entropy $\mathbb{H}(S, \mathbf{B}|\tilde{W}_T) = \mathbb{H}(\mathbf{B}|\tilde{W}_T) + \mathbb{H}(S|\tilde{W}_T, \mathbf{B})$.

We claim that

$$H(S|\tilde{W}_T, \mathbf{B}) \leq n\mathbb{E}[(d - \|\mathbf{B}\|_0)].$$

The reason is by conditioning on \mathbf{B} , we know the exact values for the bad coordinates in S . Therefore, the cardinality of the possible values for each data-point, conditioned on \mathbf{B} , cannot be more than $2^{d-\|\mathbf{B}\|_0}$. Thus,

$$\begin{aligned} I(\tilde{W}_T; S) &\geq H(S) - H(\mathbf{B}|\tilde{W}_T) - n(d - \mathbb{E}[\|\mathbf{B}\|_0]) \\ &= n\mathbb{E}[\|\mathbf{B}\|_0] - H(\mathbf{B}|\tilde{W}_T), \end{aligned}$$

where the last line follows from $H(S) = nd$ because each element in S is drawn i.i.d. Also, note that $\mathbb{E}[\|\mathbf{B}\|_0] = \mathbb{E}[\sum_{i=1}^d \mathbf{B}(i)] = d\mathbb{E}[\mathbf{B}(1)] = d2^{-n}$ where in the last line we used the fact that each element of S is i.i.d., and each column is a bad coordinate with probability 2^{-n} . Also, with the similar reasoning we obtain $H(\mathbf{B}) = H(\mathbf{B}(1), \dots, \mathbf{B}(d)) = dH_b(2^{-n})$, where for $x \in [0, 1]$ $H_b(x) = -x \log(x) - (1-x) \log(1-x)$ is the binary entropy function.

Then, we invoke a version of Fano's inequality, provided in Lemma D.6.4, to obtain

$$H(\mathbf{B}|\tilde{W}_T) \leq 1 + P_e H(\mathbf{B})$$

where $P_e = \inf_{M: \mathcal{W} \rightarrow \{0,1\}^d} \mathbb{P}(M(\tilde{W}_T) \neq \mathbf{B})$. Using the well-known inequality $H_b(x) \leq -x \log(x) + x$, we obtain

$$H(\mathbf{B}) \leq d(n2^{-n} + 2^{-n}) = d(n+1)2^{-n}.$$

Therefore,

$$\begin{aligned} I(\tilde{W}_T; S) &\geq nd2^{-n} - (n+1)d2^{-n}P_e - 1 \\ &= nd2^{-n} \left(1 - \frac{n+1}{n}P_e\right) - 1 \\ &\geq 1.5n^3(1 - 2P_e) - 1, \end{aligned} \tag{D.20}$$

where the last line follows from setting $d = 0.75T2^n$, $T = 2n^2$, and $(n+1)/n \leq 2$.

Next, we design an estimator Ψ to decode \mathbf{B} from \tilde{W}_T and analyze its probability

of error. Let $h = (\eta + \eta\lambda T)/2$. Then, the proposed estimator is given by

$$\Psi(w)(i) = \begin{cases} 1 & \text{if } |w(i)| \geq h \\ 0 & \text{if } |w(i)| < h \end{cases} \quad (\text{D.21})$$

for $i \in [d]$. In words, it compares each coordinate of w with a given threshold, and if it is larger than h , then that coordinate declares as a bad coordinate.

Let $V_T = W_T + \xi$. Then,

$$\begin{aligned} \mathbb{P}_e &\leq \mathbb{P}(\exists i \in [d] \text{ s.t. } \Psi(\tilde{W}_T)(i) \neq \mathbf{B}(i)) \\ &\leq \mathbb{P}(\{\exists i \in [d] \text{ s.t. } \Psi(\tilde{W}_T)(i) \neq \mathbf{B}(i)\} \cap \{\|V_T\| \leq 1\}) + \mathbb{P}(\|V_T\| \geq 1) \end{aligned} \quad (\text{D.22})$$

First we show that $\mathbb{P}(\|V_T\| \geq 1)$ is sufficiently small. From Eq. (D.8), we have $\|V_T\| \geq 1 = \|W_T\|^2 + r^2 + 2\|W_T\|r\theta(1)$. Then, as shown in Appendix D.3.3 given that $\{r \leq r_{\max} = 1 - \|W_T\|\}$, then $\|V_T\| \leq 1$. Using this we obtain

$$\begin{aligned} \mathbb{P}(\|V_T\| \geq 1) &= \mathbb{P}(\|W_T\|^2 + r^2 + 2\|W_T\|r\theta(1) \geq 1) \\ &\leq \mathbb{P}(r \geq 1 - \|W_T\|). \end{aligned}$$

Here $\xi = r\theta$ where $r = \|\xi\|$ and $\theta = \xi/\|\xi\|$. Recall from Lemma D.3.2 that under the event $\{T/2 \leq \|\mathbf{B}\|_0 \leq T\}$, $1/(2\sqrt{n}) \leq \|W_T\| \leq 1/\sqrt{n}$. Therefore,

$$\begin{aligned} \mathbb{P}(r \geq 1 - \|W_T\|) &\leq \mathbb{E} [\mathbb{P}^S[r \geq 1 - \|W_T\|] \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T]] + 1 - \mathbb{P}(T/2 \leq \|\mathbf{B}\|_0 \leq T) \\ &\leq \mathbb{E} [\mathbb{P}^S[r \geq 1 - 1/(2\sqrt{n})] \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T]] + 1 - \mathbb{P}(T/2 \leq \|\mathbf{B}\|_0 \leq T) \\ &\leq \mathbb{E} [\mathbb{P}^S[r \geq 1 - 1/(2\sqrt{n})] \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T]] + 2 \exp(-T/36), \end{aligned}$$

where the last line follows from Eq. (D.2). Observe that

$$\{r \geq 1 - 1/(2\sqrt{n})\} \subseteq \{r \geq 2\beta^*\},$$

due to $\beta^* = 0.1$. Also as $\sigma \leq \beta^*/\sqrt{d}$, we have

$$\mathbb{P}(r \geq 2\beta^*) \leq \mathbb{P}(r \geq \sqrt{4d(\sigma^*)^2}) \leq 2 \exp\left(-\frac{9d}{16}\right),$$

where the last inequality comes from the concentration bounds for r in Corollary D.6.3.

Since $r \perp\!\!\!\perp S$, we have

$$\mathbb{P}^S[r \geq 1 - 1/(2\sqrt{n})] \leq 2 \exp\left(-\frac{9d}{16}\right).$$

Therefore,

$$\mathbb{P}(\|V_T\| \geq 1) \leq 2 \exp\left(-\frac{9d}{16}\right) + 2 \exp(-T/36). \quad (\text{D.23})$$

Since under the event $\|V_T\| \leq 1$, $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T) = W_T + \xi$,

$$\begin{aligned} & \mathbb{P}(\{\exists i \in [d] \text{ s.t. } \Psi(\tilde{W}_T)(i) \neq \mathbf{B}(i)\} \cap \{\|V_T\| \leq 1\}) \\ &= \mathbb{P}(\{\exists i \in [d] \text{ s.t. } \Psi(W_T + \xi)(i) \neq \mathbf{B}(i)\} \cap \{\|V_T\| \leq 1\}) \\ &\leq \mathbb{P}(\{\exists i \in [d] \text{ s.t. } \Psi(W_T + \xi)(i) \neq \mathbf{B}(i)\}). \end{aligned} \quad (\text{D.24})$$

By the definition of the error probability

$$\mathbb{P}(\forall i \in [d] \Psi(W_T + \xi)(i) = \mathbf{B}(i)) \geq \mathbb{E}[\mathbb{P}^S[\forall i \in [d] \Psi(W_T + \xi)(i) = \mathbf{B}(i)] \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T]]. \quad (\text{D.25})$$

Note that W_T and \mathbf{B} are S -measurable. Therefore, the inner probability is only over ξ . Also, let $\mathcal{B} = \{i_1, \dots, i_{\|\mathbf{B}\|_0}\}$ denote the set of bad coordinates. Using the closed-form expression in Lemma D.3.1 for W_T under the event $\{T/2 \leq \|\mathbf{B}\| \leq T\}$, we have

$$\mathbb{P}^S[\forall i \in [d] \Psi(W_T + \xi)(i) = \mathbf{B}(i)] = \left(\prod_{i \in \mathcal{B}} \mathbb{P}^S[\eta + \xi(i) \geq h]\right) \prod_{i \in [d] \setminus \mathcal{B}} \left(\mathbb{P}^S[-W_T(i) + \xi(i) \leq h]\right). \quad (\text{D.26})$$

This identity follows from $\xi \perp\!\!\!\perp S$ and each coordinate of ξ are i.i.d. As shown in Lemma D.3.1, under the event $\{T/2 \leq \|\mathbf{B}\| \leq T\}$, $0 \leq -W_T(i) \leq \lambda\eta T$; therefore, $-W_T(i) < h$ for $i \in [d] \setminus \mathcal{B}$. We can simplify Eq. (D.26) as

$$\begin{aligned} & \left(\prod_{i \in \mathcal{B}} \mathbb{P}^S[\eta + \xi(i) \geq h]\right) \prod_{i \in [d] \setminus \mathcal{B}} \left(\mathbb{P}^S[-W_T(i) + \xi(i) \leq h]\right) \\ &= \left(1 - Q\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right)\right)^{\|\mathcal{B}\|_0} \prod_{i \in [d] \setminus \mathcal{B}} \left(1 - Q\left(\frac{h + W_T(i)}{\sigma^*}\right)\right) \end{aligned}$$

where for $x \in \mathbb{R}$, $Q(x) = \frac{1}{\sqrt{2\pi}} \int_{t \geq x} \exp(-\frac{t^2}{2}) dt$ is the tail distribution function of the Gaussian distribution with mean zero and variance one. Since $Q\left(\frac{h + W_T(i)}{\sigma^*}\right) \leq$

$Q\left(\frac{h-\eta\lambda T}{\sigma^*}\right) = Q\left(\frac{\eta-\eta\lambda T}{2\sigma^*}\right)$ for all $i \in [d] \setminus \mathcal{B}$, we can further lower bound as

$$\begin{aligned} \mathbb{P}^S[\forall i \in [d] \Psi(W_T + \xi)(i) = \mathbf{B}(i)] &\geq \left(\prod_{i \in \mathcal{B}} \mathbb{P}^S[\eta + \xi(i) \geq h]\right) \prod_{i \in [d] \setminus \mathcal{B}} \left(\mathbb{P}^S[\eta\lambda T + \xi(i) \leq h]\right) \\ &\geq \left(1 - Q\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right)\right)^{\|\mathbf{B}\|_0} \left(1 - Q\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right)\right)^{d - \|\mathbf{B}\|_0}, \end{aligned} \quad (\text{D.27})$$

More precisely, since $\eta\lambda T < h < \eta$, we have $\mathbb{P}^S[\eta + \xi(i) \geq h] = \mathbb{P}^S[\xi(i) \geq h - \eta] = 1 - \mathbb{P}^S[\xi(i) \geq \eta - h]$ and $\mathbb{P}^S[\eta\lambda T + \xi(i) \leq h] = \mathbb{P}^S[\xi(i) \leq h - \eta\lambda T] = 1 - \mathbb{P}^S[\xi(i) \geq h - \eta\lambda T]$.

Therefore, we can use Eq. (D.25), Eq. (D.26), and Eq. (D.27) to obtain

$$\begin{aligned} \mathbb{P}(\forall i \in [d] \Psi(W_T + \xi)(i) = \mathbf{B}(i)) &\geq \left(1 - Q\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right)\right)^d \mathbb{P}(T/2 \leq \|\mathbf{B}\|_0 \leq T) \\ &\geq \left(1 - Q\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right)\right)^d \left(1 - 2 \exp\left(-\frac{T}{36}\right)\right), \end{aligned}$$

where in the last line we have used Eq. (D.2). Note that $\eta - \eta\lambda T \geq 0$ since $\lambda \in \mathcal{O}(1/(n\sqrt{d}))$. We can use the well-known inequality $(1 - x)^n \geq 1 - nx$ for $x \leq 1$, $n \in \mathbb{N}$ to obtain

$$\begin{aligned} 1 - \mathbb{P}(\forall i \in [d] \Psi(W_T + \xi)(i) = \mathbf{B}(i)) &\leq dQ\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right) + 2 \exp\left(-\frac{T}{36}\right) - 2dQ\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right) \exp\left(-\frac{T}{36}\right) \\ &\leq dQ\left(\frac{\eta - \eta\lambda T}{2\sigma^*}\right) + 2 \exp\left(-\frac{T}{36}\right), \end{aligned}$$

Then, we invoke the inequality $Q(x) \leq \frac{1}{2} \exp(-\frac{x^2}{2})$ for $x \geq 0$ [Wai19, Ex.2.2], to further upper bound the last equation as follows:

$$1 - \mathbb{P}(\forall i \in [d] \Psi(W_T + \xi)(i) = \mathbf{B}(i)) \leq \frac{d}{2} \exp\left(-\frac{d(\eta - \eta\lambda T)^2}{2(\beta^*)^2}\right) + 2 \exp\left(-\frac{T}{36}\right). \quad (\text{D.28})$$

Finally, by combining Eq. (D.22), Eq. (D.23), Eq. (D.24), and Eq. (D.28), we obtain

$$\mathbf{P}_e \leq \frac{d}{2} \exp\left(-\frac{d(\eta - \eta\lambda T)^2}{2(\beta^*)^2}\right) + 4 \exp\left(-\frac{T}{36}\right) + 2 \exp\left(-\frac{9d}{16}\right).$$

By setting the parameters and some simple manipulations, we obtain

$$P_e \leq n^2 2^n \exp(-2^n/n) + 6 \exp(-n^2/18). \quad (\text{D.29})$$

In Fig. D.1, we plot the upper bound in Eq. (D.29). As can be seen the upper bound is decreasing and smaller than 0.1 for $n \geq 10$.

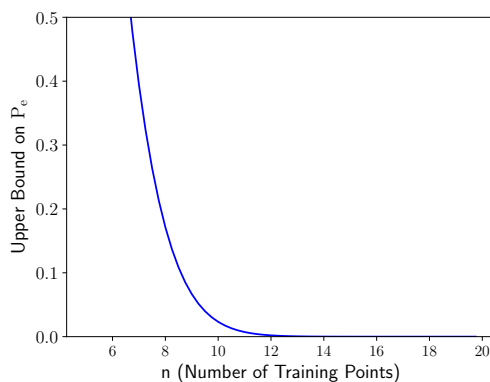


Figure D.1: The upper bound in Eq. (D.29)

Finally, combining Eq. (D.29) with Eq. (D.20), we conclude that for $n \geq 10$ if $\sigma \leq \beta^*/\sqrt{d}$, we have

$$I(\tilde{W}_T; S) \geq 1.2n^3 - 1,$$

which was to be shown.

D.3.5 Noise with Small Variance Fails: CMI

In this part of the proof we aim to show that if the variance of the noise is smaller than $\frac{(\beta^*)^2}{d}$, then $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ grows linearly with n . We begin this part of the proof with a key lemma.

We recall the definition of bad coordinates. For $i \in [d]$, we say the i -th coordinate is a *bad coordinate* iff for all $j \in [n]$, $Z_j(i) = 0$. In words, if i -th coordinate is a bad coordinate then all the entries in the i -th column of S is zero. Let $\mathbf{B} \in \{0, 1\}^d$ denote a vector such that $\mathbf{B}(i) = 1$ if and only if i is a bad coordinate. Also $\|\mathbf{B}\|_0$ denotes the number of bad coordinates.

Next, we provide a result which shows that U can be identified with high accuracy by having access to the supersample and bad coordinates. The intuition behind the result is as follows. Consider a decision making problem where by having access to \mathbf{B}

and matrix of \tilde{S} , we want to find which subset of the supersample contained in the training set, i.e., find U . First, note that, by definition, in each column of \tilde{S} exactly one sample is chosen for the training set. Also, by the definition of the bad coordinates, we know that if $i \in [d]$ is a bad coordinate, then for all $Z \in S$, we have $Z(i) = 0$. In the next theorem we show that the *uncertainty* about U is small conditioned on \mathbf{B}, \tilde{S} . The idea of the proof is to show that by only considering the bad coordinates we can *distinguish* between the points in each column of the supersample.

Lemma D.3.3. $H(U|\mathbf{B}, \tilde{S}) \leq n\mathbb{E}[2^{-\|\mathbf{B}\|_0}]$.

Proof. Let $\mathcal{B} = \{i_1, \dots, i_{\|\mathbf{B}\|_0}\} \subseteq [d]$ contains the *ordered* set of bad coordinates. For every $k \in [n]$, define the following indicator random variable

$$J_k = \mathbb{1}[\exists i \in \mathcal{B} \text{ s.t. } \tilde{Z}_{0,k}(i) \neq \tilde{Z}_{1,k}(i)]$$

Let $J = (J_1, \dots, J_n) \in \{0, 1\}^n$. Note that J is (\tilde{S}, \mathbf{B}) -measurable.

The main observation here is that provided that $J_k = 1$, then we can perfectly recover U_k . The reason is as follows: in each column of \tilde{S} , exactly one sample is a member of the training set. Also, since we know \mathbf{B} , the values of the bad coordinates are known for the points in the training set by the definition of bad coordinates. Therefore, $J_k = 1$ iff one of the point in the k -th column of \tilde{S} does not have zero on the indices in \mathcal{B} , which reveals the sample that is not in the training set. Therefore, as J is (\tilde{S}, \mathbf{B}) -measurable, we can write

$$\begin{aligned} H(U|\mathbf{B}, \tilde{S}) &= H(U|\mathbf{B}, \tilde{S}, J) \\ &= H((U)_{\{i|J_i=0\}}, (U)_{\{i|J_i=1\}}|\mathbf{B}, \tilde{S}, J) \\ &= H((U)_{\{i|J_i=0\}}|\mathbf{B}, \tilde{S}, J), \end{aligned}$$

where the last line follows from $(U)_{\{i|J_i=1\}}$ being known from J . Since the cardinality of the support of $(U)_{\{i|J_i=0\}}$ is no more than $2^{n-\|J\|_0}$, we obtain

$$H(U|\mathbf{B}, \tilde{S}) \leq n - \mathbb{E}[\|J\|_0].$$

Then, we claim that

$$\mathbb{P}(J_k = 1) = \mathbb{E}[1 - 2^{-\|\mathbf{B}\|_0}].$$

This claim conclude the proof since $\mathbb{E}[\|J\|_0] = \sum_{k=1}^n \mathbb{E}[J_k] = \sum_{k=1}^n \mathbb{P}(J_k = 1)$.

To prove the claim: $J_k = 0$ iff, conditioned on the U and \mathbf{B} , for all j such that $\mathbf{B}_j = 1$, $Z_{1-U_k,k}(j) = 0$. By the definition of the supersample, the points in the supersample are i.i.d., independent of U , and drawn from $\text{Ber}(1/2)$. Hence,

$$\mathbb{P}(J_k = 0) = \mathbb{E}[\mathbb{P}^{U,S}[J_k = 0]] = \mathbb{E}[2^{-\|\mathbf{B}\|_0}].$$

□

By the definition of mutual information, we have

$$\begin{aligned} \text{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= \mathbb{H}(U|\tilde{S}) - \mathbb{H}(U|\tilde{W}_T, \tilde{S}) \\ &= \mathbb{H}(U) - \mathbb{H}(U|\tilde{W}_T, \tilde{S}) \\ &= n - \mathbb{H}(U|\tilde{W}_T, \tilde{S}), \end{aligned} \tag{D.30}$$

where the second and third steps follow from $U \perp\!\!\!\perp \tilde{S}$ and $\mathbb{H}(U) = n$, respectively. To analyze the second term in Eq. (D.30), consider the following equality which comes from the chain rule:

$$\begin{aligned} \mathbb{H}(U, \mathbf{B}|\tilde{W}_T, \tilde{S}) &= \mathbb{H}(U|\tilde{W}_T, \tilde{S}) + \mathbb{H}(\mathbf{B}|U, \tilde{W}_T, \tilde{S}) \\ &= \mathbb{H}(\mathbf{B}|\tilde{W}_T, \tilde{S}) + \mathbb{H}(U|\tilde{W}_T, \tilde{S}, \mathbf{B}). \end{aligned}$$

Notice that $\mathbb{H}(\mathbf{B}|U, \tilde{W}_T, \tilde{S}) = 0$ as \mathbf{B} is (U, \tilde{S}) -measurable. Therefore,

$$\mathbb{H}(U|\tilde{W}_T, \tilde{S}) = \mathbb{H}(\mathbf{B}|\tilde{W}_T, \tilde{S}) + \mathbb{H}(U|\tilde{W}_T, \tilde{S}, \mathbf{B}). \tag{D.31}$$

To analyze the first term, note that conditioning cannot increase the entropy. Therefore, we have $\mathbb{H}(\mathbf{B}|\tilde{W}_T, \tilde{S}) \leq \mathbb{H}(\mathbf{B}|\tilde{W}_T)$. Then, we invoke the Fano's inequality from Lemma D.6.4 to obtain

$$\mathbb{H}(\mathbf{B}|\tilde{W}_T) \leq 1 + \mathbf{P}_e \mathbb{H}(\mathbf{B}).$$

Here, $\mathbf{P}_e = \inf_{M: \mathcal{W} \rightarrow \{0,1\}^d} \mathbb{P}(M(\tilde{W}_T) \neq \mathbf{B})$. Consider the estimator Ψ proposed in Eq. (D.21). We analyzed its probability of error in Appendix D.3.4 and obtained in Eq. (D.29) that

$$\mathbf{P}_e \leq n^2 2^n \exp(-2^n/n) + 6 \exp(-n^2/18).$$

Note that $H(\mathbf{B}) \leq d(n+1)2^{-n} \leq 2n^3$ for $n \geq 3$ as shown in Appendix D.3.4. Therefore,

$$H(\mathbf{B}|\tilde{W}_T, \tilde{S}) \leq H(\mathbf{B}|\tilde{W}_T) \leq 2n^3(n^2 2^n \exp(-2^n/n) + 6 \exp(-n^2/18)) + 1. \quad (\text{D.32})$$

Next, we analyze the second term in Eq. (D.31). Using Lemma D.3.3 we have

$$H(U|\tilde{W}_T, \tilde{S}, \mathbf{B}) \leq H(U|\tilde{S}, \mathbf{B}) \leq n\mathbb{E}[2^{-\|\mathbf{B}\|_0}]. \quad (\text{D.33})$$

The, consider

$$\mathbb{E}[2^{-\|\mathbf{B}\|_0}] = \mathbb{E}[2^{-\|\mathbf{B}\|_0} \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T]] + \mathbb{E}[2^{-\|\mathbf{B}\|_0} (\mathbf{1}[\|\mathbf{B}\|_0 < T/2] + \mathbf{1}[\|\mathbf{B}\|_0 > T])].$$

The second term can be upper bounded by $\mathbb{P}(\{\|\mathbf{B}\|_0 < T/2\} \cup \{\|\mathbf{B}\|_0 > T\})$, and this probability is less than $2 \exp(-T/36)$ as shown in Eq. (D.2). By simply upper bounding the first term by the worst-case realization, we can write

$$\begin{aligned} \mathbb{E}[2^{-\|\mathbf{B}\|_0}] &\leq \mathbb{E}[2^{-T/2} \mathbf{1}[T/2 \leq \|\mathbf{B}\|_0 \leq T]] + \mathbb{P}(\{\|\mathbf{B}\|_0 < T/2\} \cup \{\|\mathbf{B}\|_0 > T\}) \\ &\leq 2^{-T/2} + 2 \exp(-T/36). \end{aligned} \quad (\text{D.34})$$

Finally, by Eq. (D.33) and Eq. (D.34), we obtain

$$H(U|\tilde{W}_T, \tilde{S}, \mathbf{B}) \leq n(2^{-T/2} + 2 \exp(-T/36)). \quad (\text{D.35})$$

The last step is combining Eq. (D.31), Eq. (D.32), and Eq. (D.35) to lower bound $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ as

$$\begin{aligned} \text{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= n - H(U|\tilde{W}_T, \tilde{S}) \\ &\geq n - \left[n2^{-n^2} + n \exp(-n^2/18) + 2n^5 2^n \exp(-2^n/n) + 12n^3 \exp(-n^2/18) + 1 \right] \end{aligned} \quad (\text{D.36})$$

Fig. D.2 shows the upper bound on $n - \text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ in Eq. (D.36) as a function of n . As seen for $n \geq 16$, $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq n - 1.1$, and the lower bound on $\text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ is increasing.

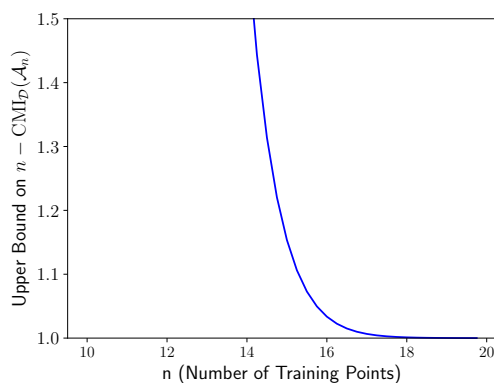


Figure D.2: Upper bound on $n - \text{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ in Eq. (D.36).

Hence, we obtain

$$\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega(n),$$

which was to be shown.

D.4 Proof of Theorem 6.6.2

The construction for proving this theorem is exactly the same as in Theorem 6.5.1.

D.4.1 Lower Bound on the Residual

For the case that $\sigma^2 \leq \text{var}_{(n)}^*$, we showed in Theorem 6.5.1 that for sufficiently large n , $\mathbb{E}[\Delta_{\sigma}(W_T) + \hat{\Delta}_{\sigma}(W_T)] = M_{\text{res}} \in \Omega(1)$. Since the loss function is 4-Lipschitz and the space has radius of 1, we have

$$\Delta_{\sigma}(W_T) + \hat{\Delta}_{\sigma}(W_T) \leq 2L\|W_T - \tilde{W}\| \leq 4LR = 16 \quad \text{a.s.}$$

Then, we invoke Lemma D.6.7 with $m = \tilde{m} = 16$ and $a = M_{\text{res}}/2$ to obtain

$$\mathbb{P}(\Delta_{\sigma}(W_T) + \hat{\Delta}_{\sigma}(W_T) > M_{\text{res}}/2) \geq \frac{M_{\text{res}}}{32}. \quad (\text{D.37})$$

D.4.2 Lower Bound on the Conditional-PAC Bayes Bound

First of all, Lemma D.6.9 implies that $\mathbb{E}^S[\text{KL}(Q(S) \parallel \frac{1}{2^n} \sum_{u \in \{0,1\}^n} Q(\tilde{S}_u))]$ is bounded by n a.s. In Theorem 6.5.1 we showed that given $\sigma^2 \geq \text{var}_{(n)}^*$ for sufficiently large n ,

we have

$$\mathbb{E}[\mathbb{E}^S[\text{KL}(Q(S) \parallel \frac{1}{2^n} \sum_{u \in \{0,1\}^n} Q(\tilde{S}_u))]] = \text{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq 0.2n.$$

Then, we use Lemma D.6.7, with the following parameters: $\tilde{m} = m = n$ and $a = 0.1n$ to obtain

$$\mathbb{P}\left(\mathbb{E}^S[\text{KL}(Q(S) \parallel \frac{1}{2^n} \sum_{u \in \{0,1\}^n} Q(\tilde{S}_u))] > 0.1n\right) \geq \frac{\text{CMI}_{\mathcal{D}}(\mathcal{A}_n) - 0.1n}{n - 0.1n} \geq \frac{1}{9}. \quad (\text{D.38})$$

D.4.3 Lower Bound on the Classical PAC-Bayes Bound

For every $s \in \{0,1\}^{n \times d}$, let $Q(s)$ denote the posterior, and $P = \mathbb{E}[Q(\tilde{S})]$ denote the prior. By construction, the training set S takes all the values in $\{0,1\}^{n \times d}$ uniformly at random. Therefore, by Lemma D.6.9, we have

$$\text{KL}(Q(S) \parallel P) = \text{KL}(Q(S) \parallel \frac{1}{2^{nd}} \sum_{s \in \{0,1\}^{n \times d}} Q(s)) \leq nd \quad \text{a.s.}$$

Consider the estimator $\Psi : \mathcal{W} \rightarrow \{0,1\}^d$ in Eq. (D.21). For every $s \in \{0,1\}^d$, let $\hat{Q} : \{0,1\}^{n \times d} \rightarrow \mathcal{M}_1(\{0,1\}^d)$ be the pushforward of $Q(s)$ through Ψ . Similarly, we can define $\hat{P} \in \mathcal{M}_1(\{0,1\}^d)$ as the pushforward of P using Ψ .

By the data-processing inequality for the KL divergence [PW19], we have

$$\text{KL}(Q(S) \parallel P) \geq \text{KL}(\hat{Q}(S) \parallel \hat{P}) \quad \text{a.s.} \quad (\text{D.39})$$

We claim that $\hat{P} = \mathbb{E}[\hat{Q}(S)]$. By a slight abuse of notation, for every $b \in \{0,1\}^d$, let $\hat{P}(b)$ denote the probability assigned to b by \hat{P} . Also, for every $s \in \{0,1\}^{n \times d}$ and a (measurable) set $A \subseteq \mathcal{W}$ let $Q(s)(A)$ be the measure assigned to set A by $Q(s)$. Similarly, we can define $P(A)$. Equipped with these notations, we can write

$$\begin{aligned} \hat{P}(b) &= \int_{w \in \mathcal{W}} P(dw) \mathbb{1}[\Psi(w) = b] \\ &= \int_{w \in \mathcal{W}} \frac{1}{2^{nd}} \sum_{s \in \{0,1\}^{n \times d}} Q(s)(dw) \mathbb{1}[\Psi(w) = b] \\ &= \frac{1}{2^{nd}} \sum_{s \in \{0,1\}^{n \times d}} \int_{w \in \mathcal{W}} Q(s)(dw) \mathbb{1}[\Psi(w) = b]. \end{aligned}$$

Here, the second step is by the definition of the prior, and the last step follows from Fubini's theorem. Notice that the expression in the last step is $\mathbb{E}[\hat{Q}(S)]$ as was to be

shown.

Recall the definition of the bad coordinates. Define the *bad coordinate profile* of $s \in \{0, 1\}^{n \times d}$ as a binary vector of length d such that its i -coordinate is one if and only if i is a bad coordinate, and it is zero otherwise. For every $b \in \{0, 1\}^d$, define set

$$\mathcal{S}_b = \{s \in \{0, 1\}^{n \times d} \mid \text{bad coordinate profile of } s \text{ is } b \}.$$

By construction, each coordinate is a bad coordinate independently with probability 2^{-n} . Therefore

$$\mathbb{P}(S \in \mathcal{S}_b) = 2^{-n\|b\|_0} (1 - 2^{-n})^{d - \|b\|_0}. \quad (\text{D.40})$$

In what follows, for every $b \in \{0, 1\}^d$ that satisfies $T/2 \leq \|b\|_0 \leq T$, we provide an upper bound on $\text{KL}(\hat{Q}(s) \parallel \hat{P})$ given $s \in \mathcal{S}_b$. We can write

$$\begin{aligned} \text{KL}(\hat{Q}(s) \parallel \hat{P}) &= \text{KL}(\hat{Q}(s) \parallel \mathbb{E}[\hat{Q}(S)]) \\ &\leq \text{KL}(\hat{Q}(s) \parallel \mathbb{P}(S \in \mathcal{S}_b) \mathbb{E}^{S \in \mathcal{S}_b}[\hat{Q}(S)] + \mathbb{P}(S \notin \mathcal{S}_b) \mathbb{E}^{S \notin \mathcal{S}_b}[\hat{Q}(S)]). \end{aligned}$$

The last line follows from the law of total expectation. Then, we invoke Lemma [D.6.8](#), to obtain

$$\begin{aligned} &\text{KL}(\hat{Q}(s) \parallel \mathbb{P}(S \in \mathcal{S}_b) \mathbb{E}^{S \in \mathcal{S}_b}[\hat{Q}(S)] + \mathbb{P}(S \notin \mathcal{S}_b) \mathbb{E}^{S \notin \mathcal{S}_b}[\hat{Q}(S)]) \\ &\leq -\log(\mathbb{P}(S \in \mathcal{S}_b)) + \text{KL}(\hat{Q}(s) \parallel \mathbb{E}^{S \in \mathcal{S}_b}[\hat{Q}(S)]). \end{aligned}$$

First, we analyze $\log(\mathbb{P}(S \in \mathcal{S}_b))$. By Eq. [\(D.40\)](#), we have

$$-\log(\mathbb{P}(S \in \mathcal{S}_b)) = n\|b\|_0 + (d - \|b\|_0) \log\left(\frac{1}{1 - 2^{-n}}\right).$$

Since $T/2 \leq \|b\|_0 \leq T$, we have $n\|b\|_0 \leq nT$. Then, using the inequality $-\log(1-x) \leq \frac{x}{1-x}$ for $x \leq 1$, we obtain $-\log(1 - 2^{-n}) \leq 2^{-n}/(1 - 2^{-n})$. Therefore,

$$\begin{aligned} (d - \|b\|_0) \log\left(\frac{1}{1 - 2^{-n}}\right) &\leq d \log\left(\frac{1}{1 - 2^{-n}}\right) \\ &\leq \frac{d2^{-n}}{1 - 2^{-n}} \\ &\leq 2d2^{-n}. \end{aligned}$$

Finally setting $d = 0.75T2^n$, we obtain the following upper bound

$$-\log(\mathbb{P}(S \in \mathcal{S}_b)) \leq \frac{5}{2}nT. \quad (\text{D.41})$$

Next, we provide an upper bound on $\text{KL}(\hat{Q}(s) \parallel \mathbb{E}^{\tilde{S} \in \mathcal{S}_b}[\hat{Q}(\tilde{S})])$. In Eq. (D.26), we analyzed the error probability of the estimator Ψ conditioned on the training set. In particular, we proved that for every training set whose number of bad coordinates is between $T/2$ and T , we have almost surely

$$\begin{aligned} \mathbb{P}^S[\exists i \in [d] \Psi(W_T + \xi)(i) \neq b(i)] &\leq n^2 2^n \exp(-2^n/n) \\ &\triangleq p_{\text{error}}. \end{aligned}$$

It implies that for all $s \in \mathcal{S}_b$ with $T/2 \leq \|b\|_0 \leq T$, $\hat{Q}(s)(b) \geq 1 - p_{\text{error}}$ and $\sum_{b' \neq b} \hat{Q}(s)(b') \leq p_{\text{error}}$. For notational convenience let $\mathbb{E}^{S \in \mathcal{S}_b}[\hat{Q}(S)] \triangleq Q_b$. By the definition of the KL divergence, we can write

$$\begin{aligned} \text{KL}(\hat{Q}(s) \parallel \hat{Q}_b) &= \sum_{b' \in \{0,1\}^d} \hat{Q}(s)(b') \log \left(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')} \right) \\ &= \hat{Q}(s)(b) \log \left(\frac{\hat{Q}(s)(b)}{\hat{Q}_b(b)} \right) + \sum_{b' \in \{0,1\}^d, b' \neq b} \hat{Q}(s)(b') \log \left(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')} \right). \end{aligned} \quad (\text{D.42})$$

Since for all $s \in \mathcal{S}_b$, $\hat{Q}(s)(b) \geq 1 - p_{\text{error}}$, we have $\hat{Q}_b(b) \geq 1 - p_{\text{error}}$. Therefore, we have

$$\begin{aligned} \hat{Q}(s)(b) \log \left(\frac{\hat{Q}(s)(b)}{\hat{Q}_b(b)} \right) &\leq \hat{Q}(s)(b) \log \left(\frac{\hat{Q}(s)(b)}{1 - p_{\text{error}}} \right) \\ &\leq -\log(1 - p_{\text{error}}). \end{aligned} \quad (\text{D.43})$$

The last step follows from $0 \leq \hat{Q}(s)(b) \leq 1$. Conditioned on $S \in \mathcal{S}_b$, the distribution of the training set is uniform over the set \mathcal{S}_b . Using this observation, for every $b' \in \{0,1\}^d$, we can write

$$\begin{aligned} \log \left(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')} \right) &= \log \left(\frac{\hat{Q}(s)(b')}{\frac{1}{|\mathcal{S}_b|} \sum_{b' \in \mathcal{S}_b} \hat{Q}(s)(b')} \right) \\ &\leq \log(|\mathcal{S}_b|) \\ &\leq \log(2^{nd}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} \sum_{b' \in \{0,1\}^d, b' \neq b} \hat{Q}(s)(b') \log \left(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')} \right) &\leq nd \sum_{b' \in \{0,1\}^d, b' \neq b} \hat{Q}(s)(b') \\ &\leq nd p_{\text{error}}. \end{aligned} \quad (\text{D.44})$$

By Eq. (D.41), Eq. (D.42), Eq. (D.43), and Eq. (D.44), we obtain

$$\begin{aligned} -\log(\mathbb{P}(S \in \mathcal{S}_b)) + \text{KL}(\hat{Q}(s) \parallel \mathbb{E}^{\tilde{S} \in \mathcal{S}_b}[\hat{Q}(\tilde{S})]) &\leq \frac{5}{2}nT - \log(1 - p_{\text{error}}) + nd p_{\text{error}} \\ &\leq \frac{5}{2}nT + \frac{p_{\text{error}}(nd + 1)}{1 - p_{\text{error}}}. \end{aligned}$$

Setting the parameters, we can see that $\frac{p_{\text{error}}(nd+1)}{1-p_{\text{error}}} \leq 1$ for $n \geq 8$.

Thus, we obtain that for every $s \in \mathcal{S}_b$ such that $T/2 \leq \|b\|_0 \leq T$, we have

$$\text{KL}(\hat{Q}(s) \parallel \hat{P}) \leq \frac{5}{2}nT + 1, \quad (\text{D.45})$$

for $n \geq 8$.

Note that the upper bound in Eq. (D.45) provides a *uniform* upper bound for every $s \in \mathcal{S}_b$ such that $T/2 \leq \|b\|_0 \leq T$. Therefore, by a simple contraposition we have

$$\begin{aligned} &\{s \in \{0,1\}^{n \times d} \mid \text{the number of bad coordinates of } s \in \{T/2, \dots, T\}\} \\ &\subseteq \{s \in \{0,1\}^{n \times d} \mid \text{KL}(\hat{Q}(s) \parallel \hat{P}) \leq \frac{5}{2}nT + 1\}. \end{aligned}$$

By considering the complement of the above statement we obtain

$$\begin{aligned} &\{s \in \{0,1\}^{n \times d} \mid \text{KL}(\hat{Q}(s) \parallel \hat{P}) > \frac{5}{2}nT + 1\} \\ &\subseteq \{s \in \{0,1\}^{n \times d} \mid \text{the number of bad coordinates of } s \notin \{T/2, \dots, T\}\}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \mathbb{P}(\text{KL}(\hat{Q}(S) \parallel \hat{P}) > \frac{5}{2}nT + 1) &\leq 1 - \mathbb{P}(T/2 \leq \|B\|_0 \leq T) \\ &\leq 2 \exp(-T/36). \end{aligned}$$

Here, the line follows from Eq. (D.2).

Next, we provide a lower bound on $\mathbb{E}[\text{KL}(\hat{Q}(S) \parallel \hat{P})]$. Let random variable \mathbf{B} denote the bad coordinate profile of S . Notice that $\mathbb{E}[\text{KL}(\hat{Q}(S) \parallel \hat{P})] = I(S; \hat{\mathbf{B}})$ where $\hat{\mathbf{B}}$ is the estimate of \mathbf{B} using the estimator Ψ . We have $I(\hat{\mathbf{B}}; S) = H(S) - H(S|\hat{\mathbf{B}})$. By construction, $H(S) = nd$. Since \mathbf{B} is a function of S , we have $H(S, \mathbf{B}|\hat{\mathbf{B}}) = H(S|\hat{\mathbf{B}})$. Then, by the chain rule for the entropy we can write $H(S, \mathbf{B}|\hat{\mathbf{B}}) = H(\mathbf{B}|\hat{\mathbf{B}}) + H(S|\mathbf{B}, \hat{\mathbf{B}})$. By conditioning on \mathbf{B} , we know the exact values for the bad coordinates in S . Therefore, the cardinality of the possible values for each data-point, conditioned on \mathbf{B} , cannot be more than $2^{d-\|\mathbf{B}\|_0}$. Therefore, we have $H(S|\mathbf{B}, \hat{\mathbf{B}}) \leq n(d - \mathbb{E}[\|\mathbf{B}\|_0])$ which gives us $H(\mathbf{B}|\hat{\mathbf{B}}) + H(S|\mathbf{B}, \hat{\mathbf{B}}) \leq H(\mathbf{B}|\hat{\mathbf{B}}) + n(d - \mathbb{E}[\|\mathbf{B}\|_0])$. By Fano's inequality in Lemma D.6.4, we have $H(\mathbf{B}|\hat{\mathbf{B}}) \leq 1 + \mathbb{P}(\hat{\mathbf{B}} \neq \mathbf{B})H(\mathbf{B})$. Therefore, we obtain

$$\begin{aligned} \mathbb{E}[\text{KL}(\hat{Q}(S) \parallel \hat{P})] &= I(S; \hat{\mathbf{B}}) \\ &\geq n\mathbb{E}[\|\mathbf{B}\|_0] - 1 - \mathbb{P}(\hat{\mathbf{B}} \neq \mathbf{B})H(\mathbf{B}) \\ &\geq nd2^{-n} - (n+1)d2^{-n}\mathbb{P}(\hat{\mathbf{B}} \neq \mathbf{B}) - 1 \\ &\geq 1.5n^3(1 - 2\mathbb{P}(\hat{\mathbf{B}} \neq \mathbf{B})) - 1. \end{aligned}$$

Here, we used the following facts. $\mathbb{E}[\|\mathbf{B}\|_0] = \mathbb{E}[\sum_{i=1}^d \mathbf{B}(i)] = d\mathbb{E}[\mathbf{B}(1)] = d2^{-n}$ since each element of S is i.i.d. and each column is a bad coordinate with probability 2^{-n} . Also, with the similar reasoning we obtain $H(\mathbf{B}) = H(\mathbf{B}(1), \dots, \mathbf{B}(d)) = dH_b(2^{-n})$, where for $x \in [0, 1]$ $H_b(x) = -x \log(x) - (1-x) \log(1-x)$ is the binary entropy function. Also, we have used the well-known inequality $H_b(x) \leq -x \log(x) + x$. Then, our analysis of the error probability of the estimator Ψ in Appendix D.3.4 implies that for $n \geq 10$, the following lower bound holds:

$$\mathbb{E}[\text{KL}(\hat{Q}(S) \parallel \hat{P})] \geq 1.2n^3 - 1.$$

In the next step, we invoke Lemma D.6.7 with the following parameters $\hat{m} = nd$, $m = \frac{5}{2}nT + 1 = 5n^3 + 1$, and $a = 0.6n^3 - 0.5$ to write

$$\begin{aligned} \mathbb{P}(\text{KL}(\hat{Q}(S) \parallel \hat{P}) \geq 0.6n^3 - 0.5) &\geq \frac{\mathbb{E}[\text{KL}(\hat{Q}(S) \parallel \hat{P})] - a - (nd - (\frac{5}{2}nT + 1))\mathbb{P}(X \geq \frac{5}{2}nT + 1)}{5n^3 + 1 - a} \\ &\geq \frac{0.6n^3 - 0.5 - 3n^3 2^n \exp(-n^2/18)}{4.4n^3 + 1.5}. \end{aligned}$$

By numerical evaluations, we can see that the lower bound is greater than 0.1 for $n \geq 16$.

From Eq. (D.39), we have

$$\begin{aligned} \mathbb{P}(\text{KL}(Q(S) \parallel P) > 0.6n^3 - 0.5) &\geq \mathbb{P}(\text{KL}(\hat{Q}(S) \parallel \hat{P}) \geq 0.6n^3 - 0.5) \\ &\geq 0.1, \end{aligned} \tag{D.46}$$

for $n \geq 16$ as was to be shown.

Concluding the Proof

In summary, in Eq. (D.37), Eq. (D.38), and Eq. (D.46), we have shown there exist constants $\alpha_1 \in \mathbb{R}_+$, $\alpha_2 \in \mathbb{R}_+$, $\alpha_3 \in \mathbb{R}_+$, $\beta_1 \in (0, 1)$, and $\beta_2 \in (0, 1)$ such that for sufficiently large n ,

1. $\mathbb{P}\left(\Delta_\sigma(W_T) + \hat{\Delta}_\sigma(W_T) > \alpha_1 \text{ or } \frac{\mathbb{E}^S[\text{KL}(Q(S) \parallel \frac{1}{2^n} \sum_{u \in \{0,1\}^n} Q(\tilde{S}_u))]}{n} > \alpha_2\right) \geq 1 - \beta_1.$
2. $\mathbb{P}\left(\Delta_\sigma(W_T) + \hat{\Delta}_\sigma(W_T) > \alpha_1 \text{ or } \frac{\text{KL}(Q(S) \parallel \mathbb{E}[Q(S)])}{n} > \alpha_3\right) \geq 1 - \beta_2.$

For notational convenience, let Bad Event₁ and Bad Event₂ denote the first and second event above.

Next, we show how this result implies the failure of PAC-Bayes bounds. Consider the decomposition of the generalization error of GD with respect to the surrogate

$$\mathbb{E}^S \left[F_{\mathcal{D}}(W_T) - \hat{F}_{S_n}(W_T) \right] \leq \mathbb{E}^S \left[F_{\mathcal{D}}(\tilde{W}) - \hat{F}_{S_n}(\tilde{W}) \right] + \hat{\Delta}_\sigma(W_T) + \Delta_\sigma(W_T),$$

Let $\text{complexity}(n)$ denote both $C_{\text{clas}}(n) \triangleq \text{KL}(Q(S) \parallel \mathbb{E}[Q(S)])$ and $C_{\text{cond}}(n) \triangleq \mathbb{E}^S[\text{KL}(Q(S) \parallel \frac{1}{2^n} \sum_{u \in \{0,1\}^n} Q(\tilde{S}_u))]$. Let $\delta < 1 - \max\{\beta_1, \beta_2\}$. Assume we instantiate the PAC-Bayes bounds with the confidence of $1 - \delta$. Then, by a simple application of the union bound we have

$$\begin{aligned} \mathbb{P}\left(\left\{ \mathbb{E}^S \left[F_{\mathcal{D}}(\tilde{W}) - \hat{F}_{S_n}(\tilde{W}) \right] \in \mathcal{O}\left(LR \sqrt{\frac{C_{\text{clas}}(n) + \log(n/\delta)}{n}}\right) \right\} \right. \\ \left. \text{and Bad Event}_1\right) &\geq 1 - \delta - \beta_1, \\ \mathbb{P}\left(\left\{ \mathbb{E}^S \left[F_{\mathcal{D}}(\tilde{W}) - \hat{F}_{S_n}(\tilde{W}) \right] \in \mathcal{O}\left(LR \sqrt{\frac{C_{\text{cond}}(n) + \log(n/\delta)}{n}}\right) \right\} \right. \\ \left. \text{and Bad Event}_2\right) &\geq 1 - \delta - \beta_2. \end{aligned}$$

Thus, we conclude that with probability at least $1 - \delta - \max\{\beta_1, \beta_2\}$ (over the

randomness in the training set) for every σ we have

$$\max\{LR\sqrt{\frac{\text{complexity}(n) + \log(n/\delta)}{n}}, \hat{\Delta}_\sigma(W_T) + \Delta_\sigma(W_T)\} \in \Omega(1),$$

as was to be shown.

D.5 Proof of Theorem 6.7.7

Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{\mathbf{e}(i) : i \in d\}$, that is, the set of all coordinate vectors in $\{0, 1\}^d$, where

$$\mathbf{e}(i) = (\underbrace{0, \dots, 0}_{i-1 \text{ times}}, 1, \underbrace{0, \dots, 0}_{d-i \text{ times}}).$$

Let the data distribution on the input be the uniform distribution, that is $\mathcal{D} = \text{Uniform}(\mathcal{Z})$. Then, we consider the simple convex, 1-Lipschitz loss function $f(w, z) = -\langle w, z \rangle$. Moreover, we consider that the weights w are in a unit ball on \mathbb{R}^d , that is $\mathcal{W} = \{w : \|w\| \leq 1\}$. Therefore, the problem is in the CLB class.

Next, we analyze the dynamics of GD. The empirical loss is given by

$$\hat{F}_{S_n}(w) = -\langle w, \hat{\mu} \rangle,$$

where $\hat{\mu}$ is the empirical mean of the instances in the training set, i.e., $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n Z_i$. Also, we have that $\partial \hat{F}_{S_n}(w) = -\hat{\mu}$ for all $w \in \mathcal{W}$. Considering the update rule of GD, i.e. $W_{t+1} = \Pi_{\mathcal{W}}(W_t + \eta \hat{\mu})$, one can show by induction that

$$W_t = \begin{cases} \eta t \hat{\mu} & \eta t \|\hat{\mu}\| \leq 1 \\ \frac{\hat{\mu}}{\|\hat{\mu}\|} & \text{Otherwise} \end{cases}. \quad (\text{D.47})$$

Now consider the \tilde{S} -measurable random variable E that is equal to one if and only if all the data instances in the supersample are distinct. That is

$$E = \mathbb{1}[\tilde{Z}_{u,i} \neq \tilde{Z}_{v,j} \text{ for all } i, j \in [n] \text{ and all } u, v \in \{0, 1\}]. \quad (\text{D.48})$$

As in the *birthday paradox problem* [MU05, Sec 5], we may bound the probability

that $E = 1$ as follows

$$\begin{aligned}\mathbb{P}(E = 1) &= \prod_{k=0}^{2n-1} \left(1 - \frac{k}{d}\right) \\ &\geq \left(1 - \frac{2n-1}{d}\right)^{2n-1},\end{aligned}\tag{D.49}$$

This way, we may engineer a dimension d for which $\mathbb{P}(E = 1) \geq c$ for all $n \geq 1$, where c is a constant probability, independent of n . Solving for Eq. (D.49) results in

$$d \geq \frac{2n-1}{1 - c^{1/(2n-1)}}.$$

For instance, for $c = 0.1$, a dimension $d = 2n^2$ suffices, and therefore $\mathbb{P}(E = 0) \leq 0.9$. Now, we are ready to study what happens to both the individual conditional mutual information $I(W_T; U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ and the evaluated mutual information $e\text{CMI}_{\mathcal{D}}(f(\text{GD}_n))$ in this particular setting.

D.5.1 Individual conditional mutual information

Note that the individual CMI may be written as follows

$$\begin{aligned}I(W_T; U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) &= H(U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) - H(U | W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) \\ &= H(U_i) - H(U_i | W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) \\ &= \log 2 - H(U_i | W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}),\end{aligned}\tag{D.50}$$

where the second and third equations follow from $U_i \perp (\tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ and $H(U_i) = \log 2$, respectively. Then, we may use Fano's inequality to bound $H(U_i | W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ and obtain the desired result. More precisely, Fano's inequality states that

$$H(U_i | W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) \leq H_b(\mathbb{P}(U_i \neq \hat{U}_i)),$$

for every estimator $\hat{U}_i(W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ and where $H_b(\cdot)$ is the binary entropy. Notice that \hat{U}_i is a function of $W_T, \tilde{Z}_{0,i}$, and $\tilde{Z}_{1,i}$. Therefore, showing that $\mathbb{P}(U_i \neq \hat{U}_i) < 0.5$ is a constant independent of n ensures that

$$I(W_T; U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) \geq \log 2 - H_b(\mathbb{P}(U_i \neq \hat{U}_i)) \in \Omega(1)\tag{D.51}$$

and completes the proof.

From Eq. (D.47), we can see that the non-zero coordinates of W_T are precisely the coordinates of the training samples. That is, if $\tilde{Z}_{U_i,i} = \mathbf{e}(k)$, then $W_T(k) \neq 0$. Therefore, under the event $E = 1$ defined in Eq. (D.48), one can precisely determine if sample $\tilde{Z}_{0,i}$ or sample $\tilde{Z}_{1,i}$ was used for training after observing W_T since the samples are all distinct. In other words, one can completely determine U_i from $(W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$. More precisely, consider a realization in which $\tilde{Z}_{0,i} = \mathbf{e}(k)$ and $\tilde{Z}_{1,i} = \mathbf{e}(l)$. Then, the estimator $\hat{U}_i(W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ is defined as $\hat{U}_i(W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) = 0$ if $W_T(k) \neq 0$ and $W_T(l) = 0$; $\hat{U}_i(W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) = 1$ if $W_T(k) = 0$ and $W_T(l) \neq 0$; otherwise in the case that $W_T(k) \neq 0$ and $W_T(l) \neq 0$, let $\hat{U}_i(W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ be a Bernoulli random variable with parameter $1/2$ independent of \tilde{S} and U . This estimator has a probability of error equal to 0 given the event $E = 1$. Therefore, the probability of error is

$$\begin{aligned} \mathbb{P}(U_i \neq \hat{U}_i) &= \mathbb{P}(E = 0) \mathbb{P}^{E=0}[U_i \neq \hat{U}_i] + \mathbb{P}(E = 1) \mathbb{P}^{E=1}[U_i \neq \hat{U}_i] \\ &= \mathbb{P}(E = 0) \mathbb{P}^{E=0}[U_i \neq \hat{U}_i] \\ &\leq 0.9 \cdot \mathbb{P}^{E=0}[U_i \neq \hat{U}_i], \end{aligned}$$

where the last line follows from the construction. Next consider the following random variables

$$G_i = \mathbb{1}[\tilde{Z}_{0,i} \neq \tilde{Z}_{1,i} \text{ and } \tilde{Z}_{1-U_i,i} \neq \tilde{Z}_{U_j,j} \text{ for all } j \neq i \in [n]],$$

which describe the situation where the given samples $\tilde{Z}_{0,i}$ and $\tilde{Z}_{1,i}$ are distinct and the sample that is not chosen is also distinct from all other samples in the dataset S , even when some of these samples are equal between themselves or to the chosen sample $\tilde{Z}_{U_i,i}$ (e.g. when $E = 0$). Therefore, given the event $E = 0$ and $G_i = 1$, the estimator \hat{U}_i still has a probability of error equal to zero. Hence, similar to before we may bound the probability of error of the estimator as

$$\begin{aligned} \mathbb{P}(U_i \neq \hat{U}_i) &\leq 0.9 \cdot \left(\mathbb{P}^{E=0}[G_i = 0] \mathbb{P}^{E=0, G_i=0}[U_i \neq \hat{U}_i] + \mathbb{P}^{E=0}[G_i = 1] \mathbb{P}^{E=0, G_i=1}[U_i \neq \hat{U}_i] \right) \\ &\leq 0.9 \cdot \mathbb{P}^{E=0, G_i=0}[U_i \neq \hat{U}_i]. \end{aligned}$$

Next, we claim that under the event where $E = 0$ and $G_i = 0$, the estimator \hat{U}_i is a Bernoulli random variable with parameter $1/2$. Consider a realization in which $U_i = u$, $\tilde{Z}_{0,i} = \mathbf{e}(k)$, and $\tilde{Z}_{1,i} = \mathbf{e}(l)$. Then, we claim that under the event $E = 0$ and $G_i = 0$, $W_T(k) \neq 0$ and $W_T(l) \neq 0$. The reason is under this event, the following cases may happen: 1) $\tilde{Z}_{0,i} = \tilde{Z}_{1,i}$ or 2) $\tilde{Z}_{0,i} \neq \tilde{Z}_{1,i}$ but there exists another sample in the training set which is equal to $\tilde{Z}_{1-u,i}$. It is easy to see that in these two cases $W_T(k) \neq 0$ and $W_T(l) \neq 0$.

Thus, we conclude that, given $E = 0$ and $G_i = 0$, we have that $\mathbb{P}^{E=0, G_i=0}[U_i \neq \hat{U}_i] = 1/2$. This is true since \hat{U}_i is a Bernoulli random variable with parameter $1/2$ independent of U and \tilde{S} . Therefore, we have that $\mathbb{P}(U_i \neq \hat{U}_i) \leq 0.45$, which completes the proof as per Eq. (D.51).

D.5.2 Evaluated conditional mutual information

Note that the evaluated CMI may be written as follows

$$\begin{aligned} e\text{CMI}_{\mathcal{D}}(f(\text{GD}_n)) &= I(F; U|\tilde{S}) \\ &= H(U|\tilde{S}) - H(U|\tilde{S}, F) \\ &= H(U) - H(U|F, \tilde{S}) \\ &= n \log 2 - H(U|F, \tilde{S}), \end{aligned} \tag{D.52}$$

where the third and fourth equations follow from $U \perp \tilde{S}$ and $H(U) = n \log 2$, respectively.

Then, as in the previous subsection, the proof relies in the fact that U can be completely determined by the loss vector F under the event $E = 1$. More precisely, note that $F_{u,i} = f(W_T, \tilde{Z}_{u,i}) = -\langle W_T, \tilde{Z}_{u,i} \rangle$. Also, remember from the previous subsection that the non-zero coordinates of W_T are precisely the non-zero coordinates of the samples that are used for training. Therefore, under the event $E = 1$, $F_{u,i} = 0$ if and only if $\tilde{Z}_{u,i}$ was not used for training and therefore $Z_i = \tilde{Z}_{1-u,i}$. Hence, one can completely determine U from F or, equivalently, $\mathbb{E} \left[H^{F, \tilde{S}, E}(U) \mathbf{1}[E = 1] \right] = 0$. We may use this fact to bound $H(U|F, \tilde{S})$ and obtain the desired result. Namely,

$$\begin{aligned} H(U|F, \tilde{S}) &= H(U|F, \tilde{S}, E) \\ &= \mathbb{E} \left[H^{F, \tilde{S}, E}(U) \mathbf{1}[E = 1] \right] + \mathbb{E} \left[H^{F, \tilde{S}, E}(U) \mathbf{1}[E = 0] \right], \\ &\leq n \cdot 0.9 \log 2 \end{aligned} \tag{D.53}$$

where the first line follows since E is \tilde{S} -measurable, and the last inequality follows from upper bounding $H^{\tilde{S}, F, G}(U)$ by $n \log 2$ and the facts that $\mathbb{E} \left[H^{F, \tilde{S}, E}(U) \mathbf{1}[E = 1] \right] = 0$ and $\mathbb{P}(E = 0) \leq 0.9$.

Finally, combining Eq. (D.52) and Eq. (D.53) results in

$$e\text{CMI}_{\mathcal{D}}(f(\text{GD}_n)) \geq n \log 2 - n \cdot 0.9 \log 2 \in \Omega(n),$$

and completes the proof.

D.6 Helper Lemmata

Lemma D.6.1 ([Ver18, Ex. 3.3.7]). *Let $X \sim \mathcal{N}(0, \mathbb{I}_d)$. Let us represent $X = R\theta$ where $R = \|X\|$ and $\theta = X/\|X\|$. Then, R and θ are independent random variables. Also, θ is uniformly distributed on the Euclidean sphere $S^{(d-1)}$ with the center at the origin.*

Lemma D.6.2 ([LM00, Lemma 1]). *Consider random vector $X \sim \mathcal{N}(0, \mathbb{I}_d)$. Then,*

$$\mathbb{P}\left(\sum_{i=1}^d a(i)X(i)^2 \geq \|a\|_1 + 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\right) \leq \exp(-t) \text{ and}$$

$$\mathbb{P}\left(\sum_{i=1}^d a(i)X(i)^2 \geq \|a\|_1 - 2\|a\|_2\sqrt{t}\right) \leq \exp(-t)$$

Corollary D.6.3. *Let $\sigma \in \mathbb{R}$, $\delta \in (0, 1)$, $d \in \mathbb{N}$, and $d \geq \log \frac{2}{\delta}$. Consider $X \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$, then*

$$\mathbb{P}\left(d\sigma^2\left(1 - 2\sqrt{\frac{\log(2/\delta)}{d}}\right) \leq \|X\|^2 \leq d\sigma^2\left(1 + 4\sqrt{\frac{\log(2/\delta)}{d}}\right)\right) \geq 1 - \delta,$$

$$\mathbb{P}(\|X\| \leq \sqrt{(1 - \alpha)d\sigma^2}) \leq 2 \exp\left(-\frac{d\alpha^2}{4}\right) \text{ for } \alpha \in [0, 1], \text{ and}$$

$$\mathbb{P}(\|X\| \geq \sqrt{(1 + \beta)d\sigma^2}) \leq 2 \exp\left(-\frac{d\beta^2}{16}\right) \text{ for } \beta \geq 0.$$

Lemma D.6.4 ([CT12, Thm. 2.10.1]). *Let X and Y be discrete random variables. Then*

$$H(X|Y) \leq H_b(P_e) + P_e H(X) \leq 1 + P_e H(X),$$

where $P_e = \mathbb{P}(\Psi(Y) \neq X)$ for any (possibly randomized) estimator Ψ of X using Y (See also [Fan52]).

Lemma D.6.5. *Let $d \in \mathbb{N}_+$. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined as $g(x) = \max\{\max_{i \in [d]} \{x(i)\}, 0\}$. Then, g is 1-Lipschitz.*

Proof. Let $x \in \mathbb{R}^d$ and $\Delta \in \mathbb{R}^d$. Let $\arg \max_{i \in [d]} \{x(i) + \Delta(i)\} = i^*$ and

$\arg \max_{i \in [d]} \{x(i)\} = j^*$ (break ties arbitrary). Then,

$$g(x + \Delta) - g(x) = \begin{cases} -x(j^*) \leq 0 & x(i^*) + \Delta(i^*) \leq 0 \text{ and } x(j^*) > 0 \\ x(i^*) + \Delta(i^*) - x(j^*) < \Delta(i^*) & x(i^*) + \Delta(i^*) > 0 \text{ and } x(j^*) > 0 \\ 0 & x(i^*) + \Delta(i^*) \leq 0 \text{ and } x(j^*) \leq 0 \\ x(i^*) + \Delta(i^*) \leq \Delta(i^*) & x(i^*) + \Delta(i^*) > 0 \text{ and } x(j^*) \leq 0 \end{cases}$$

The last case follows because $x(i^*) \leq x(j^*) \leq 0$, therefore, $x(i^*) + \Delta(i^*) \leq \Delta(i^*)$. Thus, $|g(x + \Delta) - g(x)| \leq \|\Delta\|$, as was to be shown. \square

Lemma D.6.6. *Let f be a convex and L -Lipschitz loss function, and \mathcal{W} be a convex and compact domain space with bounded diameter R . Let $\{w_t\}_{t \in [T]}$ denote the output of GD algorithm with a constant step size η . Then, we have*

$$f(w_T) - \min_{w \in \mathcal{W}} f(w) \leq \frac{R^2}{2\eta T} + \frac{(\log(T) + 2)\eta L^2}{2}$$

Proof. Let $g_t \in \partial f(w_t)$. From [Ora20, Thm. 2],

$$f(w_T) - \min_{w \in \mathcal{W}} f(w) \leq \frac{1}{T} \sum_{t=1}^T (f(w_t) - \min_{w \in \mathcal{W}} f(w)) + \frac{1}{2} \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^T \eta \|g_t\|^2.$$

Since $\|g_t\| \leq L$, the second term can be upper bounded by $\frac{\eta L^2}{2} \sum_{k=1}^{T-1} \frac{1}{k}$. Then, by the well-known bounds on the Harmonic numbers we have $\frac{\eta L^2}{2} \sum_{k=1}^{T-1} \frac{1}{k} \leq \frac{\eta L^2}{2} (\log(T-1) + 1) \leq \frac{\eta L^2}{2} (\log(T) + 1)$. For the first term, from [Bub15, Thm. 3.2], we have $\frac{1}{T} \sum_{t=1}^T (f(w_t) - \min_{w \in \mathcal{W}} f(w)) \leq \frac{R^2}{2\eta T} + \frac{\eta L^2}{2}$. Combining these two upper bounds proves the lemma. \square

Lemma D.6.7. *Let X be a random variable, $\tilde{m} \geq 0$ be a constant such that $0 \leq X \leq \tilde{m}$ a.s. Let $m \in \mathbb{R}$ be such that $0 < m \leq \tilde{m}$. Then, for every $0 \leq a < m$, we have*

$$\mathbb{P}(X > a) \geq \frac{\mathbb{E}[X] - a - (\tilde{m} - m)\mathbb{P}(X > m)}{m - a}.$$

Proof. The following holds almost surely:

$$X \leq a\mathbf{1}[X \leq a] + m\mathbf{1}[a < X \leq m] + \tilde{m}\mathbf{1}[m < X].$$

Taking an expectation concludes the proof. \square

Lemma D.6.8 ([RBS21, Lem. 2]). *Let $M \in \mathbb{N}$ and \mathcal{Y} be a measurable space. Let also $P \in \mathcal{M}_1(\mathcal{Y})$ and $Q_i \in \mathcal{M}_1(\mathcal{Y})$ for all $i \in [M]$ be probability measures. If $\alpha_i \in (0, 1)$ such that $\sum_{i=1}^M \alpha_i = 1$,*

$$\text{KL}(P \parallel \sum_{i=1}^M \alpha_i Q_i) \leq \min_{i \in [M]} \left\{ \text{KL}(P \parallel Q_i) - \log(\alpha_i) \right\}.$$

Lemma D.6.9. *Let \mathcal{Y} be a measurable space. Let $M \in \mathbb{N}$ and $P_i \in \mathcal{M}_1(\mathcal{Y})$ for $i \in [M]$ be M probability measures. Then, for every $i \in [M]$, we have*

$$\text{KL}(P_i \parallel \sum_{j=1}^M \frac{1}{M} P_j) \leq \log(M).$$

Proof. A direct application of Lemma D.6.8 gives us the result. □

Bibliography

- [VC74] V. Vapnik and A. Chervonenkis. *Theory of pattern recognition*. 1974.
- [BE02] O. Bousquet and A. Elisseeff. “Stability and generalization”. *Journal of Machine Learning Research* 2.Mar (2002), pp. 499–526.
- [D+14] C. Dwork, A. Roth, et al. “The algorithmic foundations of differential privacy”. *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.
- [RZ16] D. Russo and J. Zou. “Controlling Bias in Adaptive Data Analysis Using Information Theory”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Gretton and C. C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, 2016, pp. 1232–1240.
- [XR17] A. Xu and M. Raginsky. “Information-theoretic analysis of generalization capability of learning algorithms”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 2524–2533.
- [DHGAR21] G. K. Dziugaite, K. Hsu, W. Gharbieh, G. Arpino, and D. M. Roy. “On the role of data in PAC-Bayes bounds”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 604–612.
- [NK19] V. Nagarajan and J. Z. Kolter. “Uniform convergence may be unable to explain generalization in deep learning”. *Advances in Neural Information Processing Systems* 32 (2019).
- [BM02] P. L. Bartlett and S. Mendelson. “Rademacher and Gaussian complexities: Risk bounds and structural results”. *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

- [LW86] N. Littlestone and M. Warmuth. “Relating data compression and learnability” (1986).
- [BNSSSU16] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. “Algorithmic stability for adaptive data analysis”. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2016, pp. 1046–1059.
- [McA99a] D. A. McAllester. “Some pac-bayesian theorems”. *Machine Learning* 37.3 (1999), pp. 355–363.
- [ZBHRV17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. “Understanding deep learning requires rethinking generalization”. In: *International Conference on Representation Learning (ICLR)*. 2017. arXiv: [1611.03530v2 \[cs.LG\]](https://arxiv.org/abs/1611.03530v2).
- [LLS13] G. Lever, F. Laviolette, and J. Shawe-Taylor. “Tighter PAC-Bayes bounds through distribution-dependent priors”. *Theoretical Computer Science* 473 (2013), pp. 4–28. ISSN: 0304-3975.
- [RZ15] D. Russo and J. Zou. *How much does your data exploration overfit? Controlling bias via information usage*. 2015. arXiv: [1511.05219](https://arxiv.org/abs/1511.05219).
- [RRTWX16] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu. “Information-theoretic analysis of stability and bias of learning algorithms”. In: *2016 IEEE Information Theory Workshop (ITW)*. IEEE. 2016, pp. 26–30.
- [IEG19] I. Issa, A. R. Esposito, and M. Gastpar. “Strengthened Information-theoretic Bounds on the Generalization Error”. In: *2019 IEEE International Symposium on Information Theory (ISIT)*. 2019, pp. 582–586.
- [AAV18] A. Asadi, E. Abbe, and S. Verdú. “Chaining mutual information and tightening generalization bounds”. In: *Advances in Neural Information Processing Systems 32*. 2018.
- [BZV20a] Y. Bu, S. Zou, and V. V. Veeravalli. “Tightening Mutual Information-Based Bounds on Generalization Error”. *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 121–130.

- [SZ20a] T. Steinke and L. Zakyntinou. “Reasoning About Generalization via Conditional Mutual Information”. In: *Proceedings of the 33rd Conference On Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 3437–3452. arXiv: [2001.09122](https://arxiv.org/abs/2001.09122).
- [HGKS20] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani. “Conditioning and Processing: Techniques to Improve Information-Theoretic Generalization Bounds”. In: *Advances in Neural Information Processing Systems 34*. 2020.
- [HD20] F. Hellström and G. Durisi. “Generalization Bounds via Information Density and Conditional Information Density”. *IEEE Journal on Selected Areas in Information Theory* 1.3 (2020), pp. 824–839.
- [RBTS20] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund. “On Random Subset Generalization Error Bounds and the Stochastic Gradient Langevin Dynamics Algorithm”. In: *IEEE Information Theory Workshop (ITW)*. IEEE. 2020.
- [ZTL21] R. Zhou, C. Tian, and T. Liu. “Individually Conditional Individual Mutual Information Bound on Generalization Error”. In: *2021 IEEE International Symposium on Information Theory (ISIT)*. 2021, pp. 670–675.
- [HD21] F. Hellström and G. Durisi. “Fast-Rate Loss Bounds via Conditional Information Measures with Applications to Neural Networks”. In: *2021 IEEE International Symposium on Information Theory (ISIT)*. 2021, pp. 952–957.
- [RRT17] M. Raginsky, A. Rakhlin, and M. Telgarsky. “Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis”. In: *Proc. Conference on Learning Theory (COLT)*. 2017. arXiv: [1702.03849](https://arxiv.org/abs/1702.03849).
- [PJJ18] A. Pensia, V. Jog, and P.-L. Loh. “Generalization error bounds for noisy, iterative algorithms”. In: *2018 IEEE International Symposium on Information Theory (ISIT)*. 2018, pp. 546–550.
- [BMNSY18] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. “Learners that Use Little Information”. In: *Algorithmic Learning Theory*. 2018, pp. 25–55.

- [LM20] R. Livni and S. Moran. “A Limitation of the PAC-Bayes Framework”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 20543–20553.
- [Cat07] O. Catoni. “PAC-Bayesian supervised classification: the thermodynamics of statistical learning”. In: *Institute of Mathematical Statistics Lecture Notes-Monograph Series*. Vol. 56. 2007. arXiv: [1901.04609](#).
- [SW97] J. Shawe-Taylor and R. C. Williamson. “A PAC analysis of a Bayesian estimator”. In: *Proc. 10th Ann. Conf. Comp. Learning Theory (COLT)*. 1997, pp. 2–9.
- [McA99b] D. A. McAllester. “Some PAC-Bayesian Theorems”. *Machine Learning* 37.3 (Dec. 1999), pp. 355–363. ISSN: 1573-0565.
- [APS07] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. “Tighter PAC-Bayes bounds”. In: *Advances in Neural Information Processing Systems*. 2007, pp. 9–16.
- [PASS12] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. “PAC-Bayes bounds with data dependent priors”. *Journal of Machine Learning Research* 13.Dec (2012), pp. 3507–3531.
- [DR18a] G. K. Dziugaite and D. M. Roy. “Data-dependent PAC-Bayes priors via differential privacy”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 29. Cambridge, MA: MIT Press, 2018. arXiv: [1802.09583](#).
- [RSSPS18] O. Rivasplata, C. Szepesvari, J. S. Shawe-Taylor, E. Parrado-Hernandez, and S. Sun. “PAC-Bayes bounds for stable algorithms with instance-dependent priors”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 9214–9224. arXiv: [1806.06827](#).
- [WT11] M. Welling and Y. W. Teh. “Bayesian learning via stochastic gradient Langevin dynamics”. In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 681–688.
- [Erm75] D. L. Ermak. “A computer simulation of charged particles in solution. I. Technique and equilibrium properties”. *The Journal of Chemical Physics* 62.10 (1975), pp. 4189–4196.

- [DM17] A. Durmus and E. Moulines. “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm”. *The Annals of Applied Probability* 27.3 (2017), pp. 1551–1587.
- [MWZZ18] W. Mou, L. Wang, X. Zhai, and K. Zheng. “Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints”. In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 605–638.
- [AA19] A. R. Asadi and E. Abbe. *Chaining Meets Chain Rule: Multilevel Entropic Regularization and Training of Neural Nets*. 2019. arXiv: [1906.11148](https://arxiv.org/abs/1906.11148).
- [NHDKR19] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. “Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 11013–11023.
- [JHW17] J. Jiao, Y. Han, and T. Weissman. “Dependence measures bounding the exploration bias for general measurements”. In: *IEEE International Symposium on Information Theory*. 2017.
- [AV18] A. Lopez and V. Jog. “Generalization error bounds using Wasserstein distances”. In: *IEEE Information Theory Workshop*. 2018.
- [SSSS17] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. “Membership inference attacks against machine learning models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.
- [LLQ20] J. Li, X. Luo, and M. Qiao. “On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning”. In: *International Conference on Learning Representations*. 2020.
- [GM91] S. B. Gelfand and S. K. Mitter. “Recursive stochastic algorithms for global optimization in \mathbb{R}^d ”. *SIAM Journal on Control and Optimization* 29.5 (1991), pp. 999–1018.
- [LCB10] Y. LeCun, C. Cortes, and C. J. C. Burges. *MNIST handwritten digit database*. <http://yann.lecun.com/exdb/mnist/>. 2010.
- [Kri09] A. Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. 2009.

- [XRV17] H. Xiao, K. Rasul, and R. Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Aug. 28, 2017. arXiv: [cs.LG/1708.07747](https://arxiv.org/abs/1708.07747) [cs.LG].
- [NSY18] I. Nachum, J. Shafer, and A. Yehudayoff. “A direct sum result for the information complexity of learning”. In: *Conference On Learning Theory*. PMLR. 2018, pp. 1547–1568.
- [SZ20b] T. Steinke and L. Zakyntinou. “Open Problem: Information Complexity of VC Learning”. In: *Proceedings of the 33rd Conference On Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 3857–3863.
- [BHMZ20] O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. “Proper Learning, Helly Number, and an Optimal SVM Bound”. In: *Proceedings of the 33rd Conference On Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 582–609.
- [HK21] S. Hanneke and A. Kontorovich. “Stable Sample Compression Schemes: New Applications and an Optimal SVM Margin Bound”. In: *Algorithmic Learning Theory*. PMLR. 2021, pp. 697–721.
- [HY15] S. Hanneke and L. Yang. “Minimax analysis of active learning.” *J. Mach. Learn. Res.* 16.12 (2015), pp. 3487–3602.
- [HLW94] D. Haussler, N. Littlestone, and M. K. Warmuth. “Predicting $\{0, 1\}$ -functions on randomly drawn points”. *Information and Computation* 115.2 (1994), pp. 248–292.
- [MT07] M. Madiman and P. Tetali. “Sandwich bounds for joint entropy”. In: *2007 IEEE International Symposium on Information Theory*. IEEE. 2007, pp. 511–515.
- [HNKRD20] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite. “Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms”. In: *Advances in Neural Information Processing Systems*. 2020, pp. 9925–9935.

- [LL20] P. M. Long and R. J. Long. “On the complexity of proper distribution-free learning of linear classifiers”. In: *Algorithmic Learning Theory*. PMLR. 2020, pp. 583–591.
- [FW95] S. Floyd and M. Warmuth. “Sample compression, learnability, and the Vapnik-Chervonenkis dimension”. *Machine learning* 21.3 (1995), pp. 269–304.
- [MRT18] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [AHW87] N. Alon, D. Haussler, and E. Welzl. “Partitioning and geometric embedding of range spaces of finite Vapnik-Chervonenkis dimension”. In: *Proceedings of the third annual symposium on Computational geometry*. 1987, pp. 331–340.
- [DFHPRR15] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. “Generalization in adaptive data analysis and holdout reuse”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 2350–2358.
- [HDMR21] M. Haghifam, G. K. Dziugaite, S. Moran, and D. M. Roy. “Towards a Unified Information-Theoretic Framework for Generalization”. *Advances in Neural Information Processing Systems* 34 (2021).
- [HRVG21] H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan. “Information-theoretic generalization bounds for black-box learning algorithms”. *Advances in Neural Information Processing Systems* 34 (2021).
- [ATR21] G. Aminian, L. Toni, and M. R. Rodrigues. “Jensen-shannon information based characterization of the generalization error of learning algorithms”. In: *2020 IEEE Information Theory Workshop (ITW)*. IEEE. 2021, pp. 1–5.
- [CSDD22] E. Clerico, A. Shidani, G. Deligiannidis, and A. Doucet. “Chained Generalisation Bounds”. *arXiv preprint arXiv:2203.00977* (2022).
- [HMK22] H. Hafez-Kolahi, B. Moniri, and S. Kasaei. “Information-Theoretic Analysis of Minimax Excess Risk”. *arXiv preprint arXiv:2202.07537* (2022).

- [ZTL22] R. Zhou, C. Tian, and T. Liu. “Stochastic Chaining and Strengthened Information-Theoretic Generalization Bounds”. *arXiv preprint arXiv:2201.12192* (2022).
- [BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- [Cov69] T. M. Cover. “Learning in pattern recognition”. In: *Methodologies of pattern recognition*. Elsevier, 1969, pp. 111–132.
- [Sto76] M. Stone. “Cross-validators choice and assessment of statistical predictions (with discussion)”. *Journal of the Royal Statistical Society: Series B (Methodological)* 38.1 (1976), pp. 102–102.
- [Han16] S. Hanneke. “Refined error bounds for several learning algorithms”. *The Journal of Machine Learning Research* 17.1 (2016), pp. 4667–4721.
- [SB14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [LLS01] Y. Li, P. M. Long, and A. Srinivasan. “The one-inclusion graph algorithm is near-optimal for the prediction model of learning”. *IEEE Transactions on Information Theory* 47.3 (2001), pp. 1257–1261.
- [DDNRCWMR20] G. K. Dziugaite, A. Drouin, B. Neal, N. Rajkumar, E. Caballero, L. Wang, I. Mitliagkas, and D. M. Roy. “In Search of Robust Measures of Generalization”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020.
- [WHGC21] H. Wang, Y. Huang, R. Gao, and F. Calmon. “Analyzing the Generalization Capability of SGLD Using Properties of Gaussian Channels”. *Advances in Neural Information Processing Systems* 34 (2021).
- [SSSS09] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. “Stochastic Convex Optimization.” In: *COLT*. Vol. 2. 4. 2009, p. 5. eprint: <https://www.cs.cornell.edu/~sridharan/convex.pdf>.

- [Cau+47] A. Cauchy et al. “Méthode générale pour la résolution des systemes d’équations simultanées”. *Comp. Rend. Sci. Paris* 25.1847 (1847), pp. 536–538.
- [RM51] H. Robbins and S. Monro. “A stochastic approximation method”. *The annals of mathematical statistics* (1951), pp. 400–407.
- [Bub15] S. Bubeck. “Convex optimization: Algorithms and complexity”. *Foundations and Trends[®] in Machine Learning* 8.3-4 (2015), pp. 231–357.
- [BCLZ22] A. Banerjee, T. Chen, X. Li, and Y. Zhou. “Stability Based Generalization Bounds for Exponential Family Langevin Dynamics”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 1412–1449.
- [HMRK22] M. Haghifam, S. Moran, D. M. Roy, and G. Karolina Dziugaite. “Understanding Generalization via Leave-One-Out Conditional Mutual Information”. In: *2022 IEEE International Symposium on Information Theory (ISIT)*. 2022, pp. 2487–2492.
- [SSSS10] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. “Learnability, stability and uniform convergence”. *Journal of Machine Learning Research* 11.Oct (2010), pp. 2635–2670.
- [NDR20] J. Negrea, G. K. Dziugaite, and D. M. Roy. “In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7263–7272.
- [SGRS22] M. Sefidgaran, A. Gohari, G. Richard, and U. Simsekli. “Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 4416–4463.
- [BGZV21] Y. Bu, W. Gao, S. Zou, and V. V. Veeravalli. “Population risk improvement with model compression: an information-theoretic approach”. *Entropy* 23.10 (2021), p. 1255.

- [NDHR21] G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy. “Information-Theoretic Generalization Bounds for Stochastic Gradient Descent”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 3526–3545.
- [WM22] Z. Wang and Y. Mao. “On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications”. In: *International Conference on Learning Representations*. 2022.
- [LC02] J. Langford and R. Caruana. “(Not) Bounding the True Error”. In: *Advances in Neural Information Processing Systems 14*. Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani. MIT Press, 2002, pp. 809–816.
- [DR17] G. K. Dziugaite and D. 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. arXiv: [1703.11008](https://arxiv.org/abs/1703.11008).
- [NBS18] B. Neyshabur, S. Bhojanapalli, and N. Srebro. “A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks”. In: *International Conference on Learning Representations*. 2018.
- [ZHBHB20] M. Zhang, P. Hayes, T. Bird, R. Habib, and D. Barber. “Spread divergence”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 11106–11116.
- [CNS20] N. Chatterji, B. Neyshabur, and H. Sedghi. “The intriguing role of module criticality in the generalization of deep networks”. In: *International Conference on Learning Representations*. 2020.
- [FKMN20] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. “Sharpness-aware Minimization for Efficiently Improving Generalization”. In: *International Conference on Learning Representations*. 2020.
- [WXW20] D. Wu, S.-T. Xia, and Y. Wang. “Adversarial weight perturbation helps robust generalization”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 2958–2969.

- [PA22] A. F. Pour and H. Ashtiani. “Benefits of Additive Noise in Composing Classes with Bounded Capacity”. *arXiv preprint arXiv:2206.07199* (2022).
- [BZV20b] Y. Bu, S. Zou, and V. V. Veeravalli. “Tightening mutual information-based bounds on generalization error”. *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 121–130.
- [AKL21] I. Amir, T. Koren, and R. Livni. “SGD generalizes better than GD (and regularization doesn’t help)”. In: *Conference on Learning Theory*. 2021, pp. 63–92. arXiv: [2102.01117](https://arxiv.org/abs/2102.01117).
- [GSZ21] P. Grunwald, T. Steinke, and L. Zakyntinou. “PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes”. In: *Conference on Learning Theory*. PMLR. 2021, pp. 2217–2247.
- [HSG22] H. Harutyunyan, G. V. Steeg, and A. Galstyan. “Formal limitations of sample-wise information-theoretic generalization bounds”. *arXiv preprint arXiv:2205.06915* (2022).
- [PNG22] A. Pradeep, I. Nachum, and M. Gastpar. “Finite Littlestone Dimension Implies Finite Information Complexity”. In: *2022 IEEE International Symposium on Information Theory (ISIT)*. 2022, pp. 3055–3060.
- [NB21] M. Nokleby and A. Beirami. “Information-Theoretic Bayes Risk Lower Bounds for Realizable Models”. *arXiv preprint arXiv:2111.04579* (2021).
- [CT12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [KLMS22] T. Koren, R. Livni, Y. Mansour, and U. Sherman. “Benign Underfitting of Stochastic Gradient Descent”. *arXiv preprint arXiv:2202.13361* (2022).
- [Ora20] F. Orabona. *Last Iterate of SGD Converges (Even in Unbounded Domains)*. 2020. URL: https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/#lemmalast_average.

- [Zha04] T. Zhang. “Solving large scale linear prediction problems using stochastic gradient descent algorithms”. In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 116.
- [BFGT20] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. “Stability of stochastic gradient descent on nonsmooth convex losses”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 4381–4391.
- [SSK21] A. Sekhari, K. Sridharan, and S. Kale. “SGD: The role of implicit regularization, batch-size and multiple-epochs”. *Advances In Neural Information Processing Systems* 34 (2021), pp. 27422–27433.
- [HRS16] M. Hardt, B. Recht, and Y. Singer. “Train faster, generalize better: Stability of stochastic gradient descent”. In: *International Conference on Machine Learning*. 2016. arXiv: [1509.01240](https://arxiv.org/abs/1509.01240).
- [PW19] Y. Polyanskiy and Y. Wu. “Lecture notes on information theory”. *Lecture Notes for ECE563 (UIUC) and 6* (2019).
- [FV18] V. Feldman and J. Vondrak. “Generalization bounds for uniformly stable algorithms”. *Advances in Neural Information Processing Systems* 31 (2018).
- [FV19] V. Feldman and J. Vondrak. “High probability generalization bounds for uniformly stable algorithms with nearly optimal rate”. In: *Conference on Learning Theory*. PMLR. 2019, pp. 1270–1279.
- [RBTS21] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund. “Tighter expected generalization error bounds via Wasserstein distance”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.
- [Ora19] F. Orabona. “A modern introduction to online learning”. *arXiv preprint arXiv:1912.13213* (2019).
- [DR18b] G. K. Dziugaite and D. M. Roy. “Data-dependent PAC-Bayes priors via differential privacy”. *Advances in neural information processing systems* 31 (2018).

- [MU05] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge university press, 2005.
- [E+03] A. Elisseeff, M. Pontil, et al. “Leave-one-out error and stability of learning algorithms with applications”. *NATO science series sub series iii computer and systems sciences* 190 (2003), pp. 111–130.
- [WLF16] Y.-X. Wang, J. Lei, and S. E. Fienberg. “On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms”. In: *International Conference on Privacy in Statistical Databases*. Springer. 2016, pp. 121–134.
- [ABTRW21] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell. “An exact characterization of the generalization error for the Gibbs algorithm”. *Advances in Neural Information Processing Systems* 34 (2021), pp. 8106–8118.
- [Kem74] J. Kemperman. “On the Shannon capacity of an arbitrary channel”. In: *Indagationes Mathematicae (Proceedings)*. Vol. 77. 2. North-Holland. 1974, pp. 101–115.
- [POOAT19] B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker. “On variational bounds of mutual information” (2019). arXiv: [1905.06922](https://arxiv.org/abs/1905.06922).
- [Kal06] O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.
- [DV75] M. D. Donsker and S. S. Varadhan. “Asymptotic evaluation of certain Markov process expectations for large time, I”. *Communications on Pure and Applied Mathematics* 28.1 (1975), pp. 1–47.
- [Dur19] R. Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.
- [Te 78] H. Te Sun. “Nonnegative entropy measures of multivariate symmetric correlations”. *Information and Control* 36 (1978), pp. 133–156.
- [Han07] S. Hanneke. “Teaching dimension and the complexity of active learning”. In: *International Conference on Computational Learning Theory*. Springer. 2007, pp. 66–81.

- [EW10] R. El-Yaniv and Y. Wiener. “On the Foundations of Noise-free Selective Classification”. *Journal of Machine Learning Research* 11.5 (2010).
- [Zer08] E. Zermelo. “Untersuchungen über die Grundlagen der Mengenlehre. I”. *Mathematische Annalen* 65.2 (1908), pp. 261–281.
- [Vil09] C. Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2009.
- [Wai19] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.
- [BV04] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.
- [LM00] B. Laurent and P. Massart. “Adaptive estimation of a quadratic functional by model selection”. *Annals of Statistics* (2000), pp. 1302–1338.
- [Fan52] R. Fano. “Class notes for course 6.574: Transmission of information”. *Lecture Notes* (1952).
- [RBS21] B. Rodríguez-Gálvez, G. Bassi, and M. Skoglund. “Upper Bounds on the Generalization Error of Private Algorithms for Discrete Data”. *IEEE Transactions on Information Theory* 67.11 (2021), pp. 7362–7379.