Information Complexity of Stochastic Convex Optimization: Applications to Generalization, Memorization, and Tracing

Idan Attias [1] [2] **Gintare Karolina Dziugaite** [3] **Mahdi Haghifam** [4] **Roi Livni** [5] **Daniel M. Roy** [6] [2]

Abstract

In this work, we investigate the interplay between memorization and learning in the context of *stochastic convex optimization* (SCO). We define memorization via the information a learning algorithm reveals about its training data points. We then quantify this information using the framework of conditional mutual information (CMI) proposed by Steinke and Zakynthinou [SZ20]. Our main result is a precise characterization of the tradeoff between the accuracy of a learning algorithm and its CMI, answering an open question posed by Livni [Liv23]. We show that, in the L^2 Lipschitz–bounded setting and under strong convexity, every learner with an excess error ε has CMI bounded below by $\Omega(1/\varepsilon^2)$ and $\Omega(1/\varepsilon)$, respectively. We further demonstrate the essential role of memorization in learning problems in SCO by designing an adversary capable of accurately identifying a significant fraction of the training samples in specific SCO problems. Finally, we enumerate several implications of our results, such as a limitation of generalization bounds based on CMI and the incompressibility of samples in SCO problems.

1. Introduction

Despite intense study, the relationship between generalization and memorization in machine learning has yet to be fully characterized. Classically, ideal learning algorithms would primarily extract *relevant information* from their training data, avoiding memorization of irrelevant information. This intuition is supported by theoretical work demonstrating the benefits of limited memorization for strong generalization [LW86; RZ15; RZ16; XR17; BMNSY18; SZ20].

This intuition, however, is challenged by the success of modern overparameterized deep neural networks (DNNs). These models often achieve high test accuracy despite memorizing a significant number of training data (see, e.g., [ZBHRV17; SSSS17; CLEKS19; FZ20; CIJLT+22]). Recent studies suggest that memorization plays a more complex role in generalization than previously thought: memorization might even be *necessary* for good generalization [Fel20; FZ20; BBFST21].

In this work, we investigate the interplay between generalization and memorization in the context of *stochastic convex optimization* (SCO; [SSSS09]). A (Euclidean) SCO problem is defined by a triple $(\Theta, \mathcal{Z}, f)$, where $\Theta \subseteq \mathbb{R}^d$ is a convex subset and $f : \Theta \times \mathcal{Z} \to \mathbb{R}$ is convex in its first argument for every fixed second argument. In such an SCO problem, a learner receives a finite sample of data points in the dataspace, $\mathcal{Z}$, presumed to be drawn i.i.d. from an unknown data distribution, $\mathcal{D}$. The goal of the learner is to find an approximate minimizer of the population risk $\mathrm{F}_{\mathcal{D}}(\theta) \triangleq \mathbb{E}_{Z\sim\mathcal{D}}\left[f(\theta, Z)\right]$.

In recent years, SCO has been shown to serve as a useful theoretical model for understanding generalization in modern machine learning [Fel16; DFKL20; ACKL21; AKL21; KLMS22]. The importance of SCO can be traced to a number of factors, including: (1) it is suitable for studying gradient-based optimization algorithms, which are the workhorse behind state-of-the-art machine learning algorithms; and (2) while arbitrary empirical risk minimizers (ERMs) require sample complexity that scales with the problem dimension [Fel16; CLY23], carefully designed algorithms can achieve optimal generalization with sample complexity independent of dimension [BE02; SSSS09]. This property aligns with our goal of studying generalization in overparameterized settings such as DNNs where first-order methods output models that generalize well, despite the fact that there exist ERMs that perform poorly [ZBHRV17].

To shed light on the role of memorization in SCO, we analyze the information-theoretic properties of ε-learners for

[1] Department of Computer Science, Ben-Gurion University [2] Vector Institute [3] Google DeepMind [4] Khoury College of Computer Sciences, Northeastern University [5] Department of Electrical Engineering, Tel Aviv University. [6] Department of Statistical Sciences, University of Toronto and Vector Institute. Correspondence to: Mahdi Haghifam <haghifam.mahdi@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

SCO problems: we say a learning algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\geq 1}$ is an *ε-learner of* $(\Theta, \mathcal{Z}, f)$ if for sufficiently large n, for *every* data distribution $\mathcal{D}$, $\mathrm{F}_{\mathcal{D}}(\mathcal{A}_n(S_n)) - \min_{\theta\in\Theta} \mathrm{F}_{\mathcal{D}}(\theta) \leq \varepsilon$ with high probability over the draws of the training set $S_n \sim \mathcal{D}^{\otimes n}$ and the randomness of $\mathcal{A}$. The current paper revolves around the following fundamental question: *How much information must an ε-learner reveal about their training data?*

To address this question, we study the mutual information between (various summaries of) the learner's outputs and the training set, possibly conditional on other quantities. Early work along these lines, due to Xu and Raginsky [XR17] (see also foundational work by [RZ15; RZ16] and [NHDKR19, App. C]) provided information-theoretic generalization bounds based on the mutual information between the full training sample and the output hypothesis (the so-called *input–output mutual information*, or IOMI). Recently, Livni [Liv23] demonstrated a fundamental lowerbound on the IOMI ε-learners in the context of SCO: for every algorithm, its IOMI scales with the dimension, d. As to whether studying IOMI sheds light on memorization, there is an important caveat regarding [Liv23]: bits of information between the sample and the model do not distinguish between the number of bits per-sample and the number of memorized samples. In particular, the work of Livni [Liv23] does not rule out the sufficiency of memorizing a single example which overall has $O(d)$ entropy.

To remedy this, our work introduces a refined perspective on capturing memorization, focusing on *conditional mutual information* (CMI) as a notion of information complexity [SZ20]. CMI quantifies the amount of information that the learner's output reveals about its training sample, conditioned on a "super sample", from which the training sample is taken. (Formal definitions are provided in Section 3.) Contrasted with the bound in [XR17], in this setup, the memorization of a single example provides at most one bit of information. In other words, the scale of the CMI is more instructive on the *number* of memorized samples. Can we use CMI to fully characterize the interplay between memorization and learning in SCO?

1.1. Contributions

Our main result is a precise characterization of the tradeoff between the accuracy of a learning algorithm and its CMI:

Key result: CMI–Accuracy Tradeoff for ε-learners.

We show that in the general SCO setup as well as under further structural assumption of strong convexity, there exists a tradeoff between the accuracy of an ε-learner and its CMI: Surprisingly, to achieve small excess error, a learner *must* carry a large amount of CMI, scaling with the optimal sample size. This result completely answers an open question by Livni [Liv23]. More precisely, we study CMI of learners for two important classes of SCO problems:

- *Lipschitz bounded SCO:* We construct an SCO problem such that, for every ε-learner, there exists a distribution such that the CMI of the learner is $\Omega(1/\varepsilon^2)$, despite the already-established optimal sample complexity $O(1/\varepsilon^2)$. We complement this result with a matching upperbound. We also show that this result is true for both proper as well as improper (unconstrained) learning algorithms.
- *Strong Convexity:* Under further structural assumption of strong convexity, we establish an $\Omega(1/\varepsilon)$ lower bound on CMI of every ε-learner which we show is also tight.

Our proof techniques are inspired from the privacy literature and build on so-called fingerprinting lemmas [BUV14; Ste16; KLSU19]. Our key results and proof ideas have various interesting implications:

Limitation of the CMI Generalization Bound for SCOs. Our lower bounds highlight that CMI-based generalization bounds for SCO do not fully explain the optimal excess error. For algorithms with optimal sample complexity, the established CMI lower bound implies that standard CMI generalization guarantees are vacuous.

In more detail, Steinke and Zakynthinou [SZ20] show that the generalization error of any learner can be bounded by

$$\text{generalization gap} \leq \sqrt{\frac{\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}}.$$

(See Section 3 for a more formal statement.) Plugging our lower bound on CMI into the above equation we obtain an upper bound on the generalization gap of $O\left(\sqrt{\frac{1}{\varepsilon^2 \cdot n}}\right)$ which is strictly larger than the true $O(\varepsilon)$ error. In particular, for the optimal choice of n, we obtain a vacuous generalization bound of order $\Omega(1)$, even though the algorithm perfectly learns. Similarly, under the assumption of strong convexity, one can learn with sample complexity of $O(1/\varepsilon)$. Thus, again we obtain that the CMI bound may be order of $\Omega(1)$, even though the learner is able to learn.

Necessity of Memorization. Inspired by CMI and membership inference [CCNST+22], we have developed a framework to quantify memorization in SCO: informally, a point is considered memorized if adversary can guess correctly if this point appeared in the training set with a high confidence. Building on our construction for CMI, we design an adversary capable of correctly identifying a significant fraction of the training samples in certain SCO problems, implying that memorization is a necessary component in this context. A similar point appeared in [FV19; Fel20; BBFST21].

To be more precise, we consider a contestant and an adversary. The contestant gets to train a model on training set not revealed to the adversary. The contestant then shows the adversary a sample either from the training set, or a freshly drawn sample (not seen during training time). A point is considered *memorized* if the adversary correctly identifies whether the shown sample appeared during training time (while refraining from accusing freshly drawn samples).

We show that our approach for lower bounding CMI lets us design an adversary with the following guarantee: there exists an SCO problem such that for every ε-learner, there exists a distribution such that the adversary can distinguish $\Omega(1/\varepsilon^2)$ of the training samples with a high confidence. We also establish a similar result under an additional assumption of strong convexity, showing that there exists an adversary that can distinguish $\Omega(1/\varepsilon)$ of the training samples. Notice, that in both cases, the size of sample to be memorized scales linearly with the sample complexity. In other words, Any sample-efficient learner needs to memorize a constant fraction of its training set.

Incompressibility of Samples in SCOs. Our results rule out the existence of constant-sized (dimension-independent) sample compression schemes for SCO. Many learning algorithms, like support vector machines (SVMs), generate their output using only a small subset of training examples— for SVMs such a subset is known as support vectors. *Sample compression schemes*, introduced by Littlestone and Warmuth [LW86], provide a precise characterization of this algorithmic property. Since the optimal sample complexity in SCO is dimension-independent, a natural question to ask is whether we can construct a sample compression scheme of *constant* size for SCOs. (Here *constant compression size* refers to a dimension-independent quantity.) Using the results connecting CMI and sample compression scheme in [SZ20], we show that such a construction is impossible. This finding is in stark a contrast with binary classification [MY16].

Individual-Sample variant of CMI. We show that our techniques extend to lower-bounding the individual sample variant of CMI proposed in [HNKRD20; RBTS20; ZTL22].

1.2. Organization

The rest of this paper is is structured as follows. In Section 2 we discuss the related work. After providing the necessary preliminaries in Section 3, we present an overview of the main results in Section 4. Then, in Section 5, we discuss several implications of our main results. Finally, in Section 6 and Section 7, we present the key steps of the proofs of the main results.

2. Related Work

Information-Theoretic Measures of Generalization. In recent years, there has been a flurry of interest in the use of information-theoretic quantities for characterizing the expected generalization error of learning algorithms. For an excellent overview of recent advances see [Alq21; HDGR23]. Here, we discuss the work on worst-case information-theoretic measures of learning algorithms. The initial focus of this line of work [RZ15; RZ16; XR17] was based on *input–output mutual information (IOMI)* of an algorithm. Unfortunately, IOMI does not yield a useful notion of information complexity for learning in many key settings. For instance, prior work, in the settings of binary classification [BMNSY18; NSY18; LM20] and SCO [Liv23] highlights severe limitations of the IOMI framework: for every *good* learning algorithm in binary classification (SCO), there always exists a learning problem in which IOMI is unbounded (dimension-dependent). The notion of CMI [SZ20; GSZ21; HRVG21; HDMR21; HMRK22; HD22] remedies some of the above issues, at least in the classification setting. Despite CMI addressing some of the limitations of IOMI, Haghifam, Rodriguez-Galvez, Thobaben, Skoglund, Roy, and Dziugaite [HRTSR+23] show that CMI cannot explain the minimaxity of gradient descent in SCO. Our work significantly extends their result: we show that the same limitations hold for *every* ε-learner algorithm with a dimension-independent sample complexity. Notice that gradient descent with a proper learning rate [BFGT20; ACKL21] is one of the ε-learner algorithms that can have dimension-independent sample complexity. See Remark 5.3 for a detailed discussion. A recent work of Wang and Mao [WM23] proposes a new measure similar to CMI refereed to as hypotheses-conditioned CMI and shows that it is related to the uniform stability [BE02]. However, hypotheses-conditioned CMI is not an appropriate measure for studying memorization in SCOs since its conditioning term is different.

Memorization. [FZ20; Fel20; BBFST21; BBS22] theoretically study the necessity of memorization in learning. The measure of memorization in our work is different from the prior work. Also, the mentioned work does not study the question of memorization in the context of SCOs. Most similar to our work is [BBFST21] where the authors study memorization using IOMI. Memorization has been demonstrated to happen also empirically in state-of-the-art algorithms [CLEKS19; CTWJH+21; HVYSI22; CCNST+22]. In contrast with empirical studies, the aim of a theoretical investigation is to study its role, and whether it is *necessary* or a byproduct of current practices.

Fingerprinting Codes and Privacy Attacks. The key idea behind our lower bound proof builds on privacy attacks developed in differential privacy known as *fingerprinting* *codes* [BS95; Tar08; BUV14; Ste16; KLSU19]. Dwork, Smith, Steinke, Ullman, and Vadhan [DSSUV15] consider the problem of designing privacy attacks on the mean estimators that expose a fraction of the training data. They propose an adversary and show that every algorithm that precisely estimates mean in ℓ_∞ leaks the membership of the samples in the training set. The ℓ_∞ hypercube cannot be learned in a dimension independent sample size, therefore, to obtain the separation we desire, we can only assume a weaker ℓ_2 approximation, which leads to further challenges, especially in the unconstrained non-strongly convex case which is the hardest.

3. Preliminaries

3.1. Stochastic Convex Optimization (SCO)

A *stochastic convex optimization* (SCO) problem is a triple $(\Theta, \mathcal{Z}, f)$, where $\Theta \subseteq \mathbb{R}^d$ is a convex set and $f(\cdot, z) : \Theta \to \mathbb{R}$ is a convex function for every $z \in \mathcal{Z}$. We refer to Θ as the parameter space, to its elements as parameters, to elements of $\mathcal{Z}$ as data, and to f as the *loss function*. Informally, given an SCO problem $(\Theta, \mathcal{Z}, f)$, the goal is to find an approximate minimizer of the *population risk* $\mathrm{F}_{\mathcal{D}}(\theta) \triangleq \mathbb{E}_{Z\sim\mathcal{D}}[f(\theta, Z)]$, given an i.i.d. sample $S_n = \{Z_1, \dots, Z_n\}$ drawn from an unknown distribution $\mathcal{D}$ on $\mathcal{Z}$, denoted by $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$. The *empirical risk* of $\theta \in \Theta$ on a sample $S_n \in \mathcal{Z}^n$ is $\hat{\mathrm{F}}_{S_n}(\theta) := \frac{1}{n}\sum_{i\in[n]} f(\theta, Z_i)$, 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 $\hat{\theta} = \mathcal{A}_n(S_n)$ in $\mathbb{R}^d$. The *expected generalization error* of $\mathcal{A}_n$ under $\mathcal{D}$ is $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) = \mathbb{E}[\mathrm{F}_{\mathcal{D}}(\mathcal{A}(S_n)) - \hat{\mathrm{F}}_{S_n}(\mathcal{A}(S_n))]$. Also, the expected excess error $\mathcal{A}_n$ under $\mathcal{D}$ is $\mathbb{E}[\mathrm{F}_{\mathcal{D}}(\mathcal{A}(S_n))] - \min_{\theta\in\Theta} \mathrm{F}_{\mathcal{D}}(\theta)$. A learning algorithm is called *proper* if its output, for all possible training sets satisfies $\mathcal{A}_n(S_n) \in \Theta$. Otherwise, it is called *improper*.

Definition 3.1. (ε-learner for SCO) Fix an SCO problem $(\Theta, \mathcal{Z}, f)$ and $\varepsilon > 0$. We say $\mathcal{A} = \{\mathcal{A}_n\}_{n\geq 1}$ *ε-learns* $(\Theta, \mathcal{Z}, f)$ with sample complexity of $N : \mathbb{R} \times \mathbb{R} \to \mathbb{N}$ if the following holds: for every $\delta \in (0, 1]$, given number of samples $n \geq N(\varepsilon, \delta)$, we have that for every $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, with probability at least $1 - \delta$ over $S_n \sim \mathcal{D}^{\otimes n}$ and internal randomness of $\mathcal{A}$,

$$\mathrm{F}_{\mathcal{D}}(\mathcal{A}_n(S_n)) - \min_{\theta\in\Theta} \mathrm{F}_{\mathcal{D}}(\theta) \leq \varepsilon.$$

We also refer to $N(\cdot, \cdot)$ as *sample complexity* of $\mathcal{A}$.

We consider two important subclasses of SCO problems that impose different conditions over the loss function and the parameter space [SB14; SSSS09].

1. *Convex-Lipschitz-Bounded (CLB):* SCO with convex and L-Lipschitz loss function defined over a bounded domain with diameter R, namely, for any $\theta \in \Theta$ we have $\|\theta\| \leq R$. We say a loss function is L-Lipschitz if and only if $\forall z \in \mathcal{Z}$, $\forall \theta_1, \theta_2 \in \Theta : |f(\theta_1, z) - f(\theta_2, z)| \leq L \|\theta_2 - \theta_1\|$. We refer to this subclass as $\mathcal{C}_{L,R}$.
2. *SCO with L-Lipschitz and λ-strongly convex loss (CSL):* We say a loss function is λ-strongly convex for all $\theta_1, \theta_2 \in \Theta$ and $z \in \mathcal{Z}$ we have $f(\theta_2, z) \geq f(\theta_1, z) + \langle \partial f(\theta_1, z), \theta_2 - \theta_1\rangle + \frac{\lambda}{2}\|\theta_2 - \theta_1\|^2$ where $\partial f(\theta_1, z)$ is the subgradient of $f(\cdot, z)$ at w. The definition of Lipschitzness is the same as in the CLB subclass. We refer to this subclass as $\mathcal{C}_{L,\lambda}$.

3.2. Measure of Information Complexity

Next, we formally introduce the framework proposed by Steinke and Zakynthinou [SZ20] which aims to quantify the information complexity of a learning algorithm.

Definition 3.2. Let $\mathcal{D}$ be a data distribution, and $\mathcal{A} = (\mathcal{A}_n)_{n\geq 1}$ a learning algorithm. For every $n \in \mathbb{N}$, let $\tilde{\boldsymbol{Z}} = (Z_{i,j})_{i\in\{0,1\}, j\in[n]}$ be an array of i.i.d samples drawn from $\mathcal{D}$, and $U = (U_1, \dots, U_n) \sim \mathrm{Ber}\left(\frac{1}{2}\right)^{\otimes n}$, where U and $\tilde{\boldsymbol{Z}}$ are independent. Define a training set $S_n = (Z_{U_i,i})_{i\in[n]}$. The conditional mutual information (CMI) of $\mathcal{A}_n$ with respect to $\mathcal{D}$ is

$$\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \triangleq I(\mathcal{A}_n(S_n); U|\tilde{\boldsymbol{Z}}).$$

4. Main Results

In this section we formally state our main results. First in Section 4.1, we give an overview of CMI-accuracy tradeoff for ε-learners. Then, in Section 4.2, we precisely define the memorization game and present our results on the necessity of memorization.

4.1. CMI-Accuracy Tradeoff

We begin with a lower bound on the CMI for the CLB subclass.

Theorem 4.1. *There exists a loss function $f(\cdot, z)$ that is $O(1)$-Lipschitz, for every z such that: For every $\varepsilon \leq 1$ and for every algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ that ε-learns with the sample complexity $N(\cdot, \cdot)$ the following holds: for every $\delta \leq \varepsilon$, $n \geq N(\varepsilon, \delta)$, and $d \geq \Omega(n^4 \log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) = \Omega\left(\frac{1}{\varepsilon^2}\right)$.*

In particular, we obtain that for every algorithm, in sufficiently large dimension, there exists a problem instance where the CMI-generalization bound in [SZ20] becomes vacuous for every algorithm with sample complexity $n = O(1/\varepsilon^2)$.

Notice that the bound above is tight, namely, there exists an ε-learner with CMI at most $O(1/\varepsilon^2)$. Consider a base algorithm with the sample complexity $N(\varepsilon,\delta) = O\left(\log\left(1/\delta\right)/\varepsilon^2\right)$ (e.g. regularized ERM [BE02] or stabilized Gradient Descent [BFGT20]). Then, given $n \geq \Omega\left(\log\left(1/\delta\right)/\varepsilon^2\right)$, we may consider an algorithm that subsamples $O(\log\left(1/\delta\right)/\varepsilon^2)$ examples and feed it into the base algorithm. By the definition of the CMI, it is bounded by the size of the subsample used for learning. This argument shows that there exists an algorithm with $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) = O(1/\varepsilon^2)$. Formal statement of the described upperbound appears in Theorem 6.5.

Under further structural assumptions, though, the sample complexity in SCO can be improved. It is a question then if CMI bounds can also be further tightened under such structural assumptions such as, for example strong convexity. Our next result shows that this is indeed the case:

Theorem 4.2. *There exists a function $f(\cdot,z)$ that is $O(1)$-Lipschitz, and $O(1)$-strongly convex, for every z such that: For every $\varepsilon < 1/24$ and $\delta < 1/48$ and for every ε-learner ($\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$), with sample complexity $N(\cdot,\cdot)$ the following holds: for every $n \geq N(\varepsilon,\delta)$, $\delta < O(1/n^2)$, and $d \geq O(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega\left(\frac{1}{\varepsilon}\right)$.*

As in the general case, the above bound is tight. As discussed in [SSSS09], any ERM is stable, hence generalizes over a strongly convex objective with sample complexity of $N(\varepsilon,\delta) = O(\log(1/\delta)/\varepsilon)$. Therefore, as before, we obtain that the above bound is tight for this setup. Formal statement of the upperbound appears in Theorem 7.4.

We finish this section by introducing a memorization game that helps us formalize in what sense a learner must memorize the data in SCO.

4.2. Memorization Game

Intuitively, we can think of CMI as measuring the number of examples we can identify from the training set by observing the model. However, formally there is a gap between this interpretation and the definition of CMI. For example, one could think of a learner that *spreads the information* by using many samples, where we have that $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega(1/\varepsilon^2)$, but for each specified example, the information over U_i is small (see Definition 3.2.). In other words, there is a formal gap between large CMI and intuitive notions of memorization. In this subsection, we aim to close this gap by showing that, in fact, this is not the case, and the information the learner carries on U can be used to actually identify examples from the training set. The proofs will be appeared in Appendix F.

Definition 4.3 (Recall Game for i-th example)**.** Let $\mathcal{A} = \{\mathcal{A}_n\}_{n\geq 1}$ be a learning algorithm, $S_n = (Z_1,\dots,Z_n) \sim \mathcal{D}^{\otimes n}$ be a training set, and $\hat{\theta} = \mathcal{A}_n(S_n)$. Let $\mathcal{Q} : \mathbb{R}^d \times \mathcal{Z} \times \mathcal{M}_1(\mathcal{Z}) \to \{0,1\}$ be an adversary. Consider the following game. For $i \in [n]$, we sample a fresh data point $\tilde{Z}_i \sim \mathcal{D}$, independent of $\hat{\theta}$ and Z_i. Let $Z_{1,i} = Z_i$ and $Z_{0,i} = \tilde{Z}_i$. Then, we flip a fair coin $b_i \sim \mathrm{Unif}(\{0,1\})$. Finally, the adversary outputs $\hat{b}_i \triangleq \mathcal{Q}\left(\hat{\theta}, Z_{b_i,i}, \mathcal{D}\right)$.

The next definition formalizes the measures used for evaluating an adversary.

Definition 4.4 (soundness and recall)**.** Consider the setup described in Definition 4.3. Assume that the adversary plays the game for each of the data points in the training set, i.e., n rounds. Then,

1. We say the adversary is ξ-sound if $\mathbb{P}\left(\exists i \in [n] : \mathcal{Q}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1\right) \leq \xi$ where $\xi \in [0,1]$ is a constant.
2. We say the adversary certifies the recall of m samples if $\mathbb{P}\left(\sum_{i=1}^n \mathcal{Q}\left(\hat{\theta}, Z_{1,i}, \mathcal{D}\right) \geq m\right) \geq 1/3$.

Intuitively, soundness condition implies that if the adversary identifies a sample as part of the training set, its prediction needs to be accurate. Then, the recall condition makes sure the adversary can identify many training points, which is quantified by m. Next, we present the main results:

Theorem 4.5. *Fix $\xi \in (0,1]$. There exists a SCO problem with $O(1)$ convex Lipschitz loss defined over the ball of radius one in $\mathbb{R}^d$, and there exists an efficient adversary such that the following is true. For every $\varepsilon < 1$, $\delta < \varepsilon$, and for every ε-learner ($\mathcal{A}$), with sample complexity $N(\varepsilon,\delta) = \Theta(\log(1/\delta)/\varepsilon^2)$ the following holds: for $n = N(\varepsilon,\delta)$ and $d \geq O(n^4\log(n/\xi))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that the adversary is ξ-sound and certifies a recall of $\Omega(1/\varepsilon^2)$ samples.*

Theorem 4.6. *Fix $\xi \in (0,1]$. There exists a SCO problem with $O(1)$ strongly convex and $O(1)$ Lipschitz loss, and there exists an efficient adversary such that the following is true. For every $\varepsilon < 1/24$, $\delta < 1/48$, and for every ε-learner ($\mathcal{A}$), with sample complexity N the following holds: for every $n \geq N(\varepsilon,\delta)$, $\delta < O(1/n^2)$, and $d \geq O(n^4\log(n/\xi))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that the adversary is ξ-sound and certifies a recall of $\Omega(1/\varepsilon)$ samples.*

Remark 4.7. Notice that the adversary only requires access to the the output of the algorithm. Moreover, in Definition 4.3, we assume that the adversary has access to the data distribution. This assumption is only for convenience and can be easily relaxed by assuming the adversary has a constant number of fresh samples from the unknown data distribution. As can be seen in the proof, the adversary only requires an estimate of $\mu = \mathbb{E}_{Z\sim\mathcal{D}}[Z]$. ◁

5. Implications

5.1. Limitation of CMI-Based Generalization Bounds for SCO

CMI is proposed by Steinke and Zakynthinou [SZ20] as an information-theoretic measure for studying the generalization properties of learning algorithms. An important question regarding CMI framework is that for which learning problems and learning algorithms is the CMI framework *expressive* enough to accurately estimate the optimal worst-case generalization error? This question has been studied extensively for the the setting of binary classification and 0–1 valued loss. In [SZ20; GSZ21; HDMR21; HRVG21; HD22], the authors show that CMI framework can be used to establish optimal worst-case excess error bounds in the realizable setting. Despite these successful applications, much less is known about the optimality or limitations of CMI framework beyond the setting of binary classification and 0–1 valued loss. In this section, our main result is that for every learning algorithm for SCO with an optimal sample complexity, the generalization bound using CMI framework is vacuous. First, we start by quoting a result from [HRTSR+23] which extends the generalization bounds based on CMI to SCO problems.

Theorem 5.1 ([HRTSR+23]). *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 $(f, \Theta, \mathcal{Z}) \in \mathcal{C}_{L,R}$. Then, for every learning algorithm $\mathcal{A}_n$ such that $\mathcal{A}_n(S_n) \in \Theta$ a.s., $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq LR\sqrt{8\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)/n}$.*

Consider an SCO problem $(\Theta, \mathcal{Z}, f) \in \mathcal{C}_{L,R}$. To control the excess population error for an algorithm, a common strategy is bounding it using generalization and optimization error:

$$\begin{aligned}&\mathbb{E}\left[\mathrm{F}_{\mathcal{D}}(\mathcal{A}_n(S_n))\right] - \min_{\theta\in\Theta} \mathrm{F}_{\mathcal{D}}(\theta)\\ &\quad\leq \mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) + \mathbb{E}\big[\hat{\mathrm{F}}_{S_n}(\mathcal{A}_n(S_n)) - \min_{\theta\in\Theta} \hat{\mathrm{F}}_{S_n}(\theta)\big].\end{aligned}$$

For a proof see [HRTSR+23; BFGT20]. Since we are interested in controlling the $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ using CMI, we can use Theorem 5.1 to further upper-bound the excess error as

$$\begin{aligned}&\mathbb{E}\left[\mathrm{F}_{\mathcal{D}}(\mathcal{A}_n(S_n))\right] - \min_{\theta\in\Theta} \mathrm{F}_{\mathcal{D}}(\theta)\\ &\leq LR\sqrt{\frac{8\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)}{n}} + \mathbb{E}\big[\hat{\mathrm{F}}_{S_n}(\mathcal{A}_n(S_n)) - \min_{\theta\in\Theta} \hat{\mathrm{F}}_{S_n}(\theta)\big].\end{aligned} \tag{1}$$

It has been known for every learning algorithm that ε-learn the subclass $\mathcal{C}_{L,R}$ of SCOs, the optimal sample complexity is $\Theta\left(\left(\frac{LR}{\varepsilon}\right)^2\right)$[SSSS09]. A natural question to ask is: *Can the excess error decomposition using CMI accurately capture the worst-case excess error of optimal algorithms for SCOs?* Our next result provides a negative answer to this question.

Theorem 5.2. *For every $L \in \mathbb{R}$ and $R \in \mathbb{R}$, there exists an SCO problem $(\Theta, \mathcal{Z}, f) \in \mathcal{C}_{L,R}$ such that the following holds: for every learning algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ with sample complexity $N : \mathbb{R} \to \mathbb{N}$ such that for every $\varepsilon > 0$, $N(\varepsilon, \delta) = \tilde{\Theta}\left(\left(\frac{LR}{\varepsilon}\right)^2\right)$, there exists a data distribution such that $LR\sqrt{8\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)/n} = \Theta\left(LR\right)$, while the excess error is at most ε.*

Remark 5.3. In [HRTSR+23], the authors show that for a *particular* algorithm of Gradient Descent (GD) there exists a distribution such that, the upperbound based on CMI is vacuous. With the correct choice of learning rate GD can, with an optimal sample complexity, learn the subclass CLB of SCOs. Notice that our result in Theorem 5.2 significantly extends the limitations proved in [HRTSR+23], by showing that for *every* learning algorithm with an optimal sample complexity, the generalization bound based on CMI is *vacuous*. ◁

5.2. Non-Existence of Sample Compression Schemes

Many learning algorithms share the property that their output is constructed using a small subset of the training example. For example, in support vector machines, only the set of support vectors is needed to construct the separating hyperplane in the realizable setting. *Sample compression schemes*, proposed by Littlestone and Warmuth [LW86], provide a formal meaning for this algorithmic property. Formally, we say a learning algorithm $\mathcal{A}_n$ is a *sample compression scheme* of size $k \in \mathbb{N}$ if there exists a pair (κ, ρ) of maps such that, for all samples $s = (z_i)_{i=1}^n$ 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 the output of the algorithm, i.e., $\mathcal{A}_n(s) = \rho(\kappa(s))$. Steinke and Zakynthinou prove that for $n \geq k$, if $\mathcal{A}_n$ is a sample compression scheme (κ, ρ) of size k. Then for every $\mathcal{D}$, $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq k\log(2n)$ where $\mathcal{A}_n(\cdot) = \rho(\kappa(\cdot))$.

A natural question to ask is: *Can we learn CLB or CSL subclasses of SCOs using sample compression schemes?* In particular, we are interested in sample compression schemes in which k is independent of the dimension and n so that the algorithm has a dimension-independent sample complexity. Using the results presented in the previous sections, we provide a negative answer.

Corollary 5.4. *Let $\mathcal{P}_{cvx}^{(d)}$ be the problem instance described in Section 6.1.1. For every $\varepsilon \leq 1$ and $\delta \leq \varepsilon$ and for every algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ which is a sample compression of size k that ε-learns $\mathcal{P}_{cvx}^{(d)}$ with the sample complexity $\Theta\left(1/\varepsilon^2\right)$ the following holds: for every $n = \Theta(1/\varepsilon^2)$, and $d \geq \Omega(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that $k \geq \Omega(n)$.*

Corollary 5.5. *Let $\mathcal{P}_{scvx}^{(d)}$ be the problem instance described in Section 7.1.1. For every $\varepsilon < 1/24$ and $\delta < 1/48$ and* *for every algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ which is a sample compression of size k that ε-learns $\mathcal{P}_{scvx}^{(d)}$ with the sample complexity $\Theta\left(1/\varepsilon\right)$ the following holds: for every $n = \Theta(1/\varepsilon)$, $\delta < O(1/n^2)$, and $d \geq \Omega(n^4 \log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that $k \geq \Omega(n)$.*

5.3. Extensions to Individual Sample CMI

One drawback of CMI is for that many natural deterministic algorithm it can be $\Omega(n)$. This limitation can be attributed to the conditioning term in CMI which tends to reveal too much information. One notable approach to address this issue is the development of *individual sample CMI* (ISCMI) in [RBTS20; ZTL22]. Consider the structure introduced in Definition 3.2. Then, define

$$\mathrm{ISCMI}_{\mathcal{D}}(\mathcal{A}_n) \triangleq \sum_{i=1}^{n} I(\mathcal{A}_n(S); U_i | Z_{0,i}, Z_{1,i})$$

In [RBTS20; ZTL22], it has been shown for every learning algorithm and every data distribution $\mathrm{ISCMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$. Moreover, similar to CMI, small ISCMI implies generalization. Therefore, it is natural to ask: *Can we circumvent the lowerbounds proved for CMI by measuring the information complexity of ε-learners using* $\mathrm{ISCMI}_{\mathcal{D}}(\mathcal{A}_n)$*?* Our main result in this part provides a negative answer to this question. We show that exactly the same lowerbound stated in Theorem 4.1 and Theorem 4.2 hold for ISCMI. The proof is appeared on Appendix G.

Corollary 5.6. *Let $\mathcal{P}_{cvx}^{(d)}$ be the problem instance described in Section 6.1.1. For every $\varepsilon \leq 1$ and $\delta \leq \varepsilon$ and for every proper algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ that ε-learns $\mathcal{P}_{cvx}^{(d)}$ with the sample complexity $N(\cdot,\cdot)$ the following holds: for every $n \geq N(\varepsilon,\delta)$, and $d \geq \Omega(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that* $\mathrm{ISCMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega\left(\frac{1}{\varepsilon^2}\right)$.

Corollary 5.7. *Let $\mathcal{P}_{scvx}^{(d)}$ be the problem instance described in Section 7.1.1. For every $\varepsilon < 1/24$ and $\delta < 1/48$ and for ε-learns $\mathcal{A}$ for $\mathcal{P}_{scvx}^{(d)}$ with the sample complexity $N(\cdot,\cdot)$ the following holds: for every $n \geq N(\varepsilon,\delta)$, $\delta < O(1/n^2)$, and $d \geq O(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that* $\mathrm{ISCMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega\left(\frac{1}{\varepsilon}\right)$

6. Characterization of CMI for the CLB SCOs

In this section and Section 7, we discuss the key steps of the proof of CMI lowerbounds. We begin with a characterization of CMI of ε-learners for CLB subclasses of SCOs (All the proofs appear in Appendix C).

For the general case that we do not impose any condition on the output of the learner, the proof turns out to be slightly more subtle. In particular, there is a technical difference between proving the result for *improper (unconstrained)* learners and *proper (constrained)* learners. This issue does not appear in the strongly convex case as discussed in Remark 7.2. Therefore, we begin by first proving an intermediate result for *proper learners*.

Remark 6.1. Notice that by simply scaling the problem, we can reduce the lowerbound for $\mathcal{C}_{L,R}$ with an arbitrary L, R to $\mathcal{C}_{1,1}$. Therefore, for the rest of this section, we focus on $\mathcal{C}_{1,1}$. Also, without loss of generality, we can assume the parameter space is given by $\mathcal{B}_d(1)$. ◁

6.1. Lower Bound for Proper Learners

6.1.1. Construction of a Hard Problem Instance for proper learners

Let $d \in \mathbb{N}$. Let $\mathcal{Z} = \{\pm 1/\sqrt{d}\}^d$ and $\Theta = \mathcal{B}_d(1)$. Define the loss function $f : \Theta \times \mathcal{Z} \to \mathbb{R}$ as $f(\theta, z) = -\langle \theta, z\rangle$. It is immediate to see that $f(\cdot, z)$ is 1-Lipschitz. Let $\mathcal{P}_{\text{cvx}}^{(d)} \triangleq (\Theta, \mathcal{Z}, f)$ be the described SCO problem.

6.1.2. Properties of ε-Learners

In this section, we prove several properties that are shared between every ε-learners for $\mathcal{P}_{\text{cvx}}^{(d)}$.

Lemma 6.2. *Fix $\varepsilon > 0$. Let $\mathcal{A}$ be an ε-learner for $\mathcal{P}_{cvx}^{(d)}$ with the sample complexity of $N(\cdot,\cdot)$. Then, for every $\delta > 0$, $n \geq N(\varepsilon,\delta)$ and every $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, with probability at least $1-\delta$, we have $\|\mu\| - \varepsilon \leq \left\langle \hat{\theta}, \mu \right\rangle$, and, $\|\mu\| - \varepsilon - 2\delta \leq \mathbb{E}\left[\left\langle \hat{\theta}, \mu \right\rangle\right]$ where $\hat{\theta} = \mathcal{A}_n(S_n)$ and $\mu = \mathbb{E}_{Z\sim\mathcal{D}}\left[Z\right]$.*

The main implication of Lemma 6.2 is that the output of an accurate learner has a significant correlation to the mean of the data distribution. As the learner does not know the data distribution, in the next result we show that the correlation to the mean of an unknown data distribution translates to a correlation between the output and the samples in the training set. The construction of the data distribution is based on the techniques developed by Kamath, Li, Singhal, and Ullman [KLSU19].

Lemma 6.3. *Fix $\varepsilon > 0$. For every ε-learner $\mathcal{A}$ for $\mathcal{P}_{cvx}^{(d)}$ with sample complexity $N(\cdot,\cdot)$, there exists $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, such that for every $\delta > 0$*

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{d}\left(\frac{144\varepsilon^2 - d(\mu^{(k)})^2}{1 - d(\mu^{(k)})^2}\right)\left(\hat{\theta}^{(k)}\right)\left(Z_i^{(k)} - \mu^{(k)}\right)\right] \geq 2\varepsilon - 4\delta,$$

where $n \geq N(\varepsilon,\delta)$, $S_n = (Z_1,\ldots,Z_n) \sim \mathcal{D}^{\otimes n}$, and $\mu = \mathbb{E}_{Z\sim\mathcal{D}}[Z]$.

6.1.3. CMI-Accuracy Tradeoff for CLB

Theorem (Restatement of Theorem 4.1)**.** *Let $\mathcal{P}_{cvx}^{(d)}$ be the problem instance described in Section 6.1.1. For every*

$\varepsilon \leq 1$ and $\delta \leq \varepsilon$ and for every proper algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ that ε-learns $\mathcal{P}_{\text{cvx}}^{(d)}$ with the sample complexity $N(\cdot,\cdot)$ the following holds: for every $n \geq N(\varepsilon,\delta)$, and $d \geq \Omega(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega\left(\frac{1}{\varepsilon^2}\right)$.

Proof Sketch. Let $\mathcal{P}_{\text{cvx}}^{(d)}$ be the problem instance described in Section 6.1.1. Fix an ε-learner $\mathcal{A}$ for $\mathcal{P}_{\text{cvx}}^{(d)}$, and let the data distribution be such that it satisfies Lemma 6.3. Consider the structure introduced in the definition of CMI in Definition 3.2 and define diagonal matrix $A \in \mathbb{R}^{d\times d}$ where $A = \operatorname{diag}\left[\left\{\frac{144\varepsilon^2 - d(\mu^{(k)})^2}{1-d(\mu^{(k)})^2}\right\}_{k=1}^{d}\right]$. For every $i \in [n]$, $T_{0,i} = \left\langle \hat{\theta}, A\left(Z_{0,i}-\mu\right)\right\rangle$ and $T_{1,i} = \left\langle \hat{\theta}, A\left(Z_{1,i}-\mu\right)\right\rangle$. Let $\bar{U}_i = 1 - U_i$. Notice that $Z_{\bar{U}_i,i} \perp\!\!\!\perp \hat{\theta}$ given U_i by the definition of CMI. Then, we show that $T_{\bar{U}_i,i}$ is a sub-Gaussian random variable with variance proxy of $O(1/\sqrt{d})$. Therefore, with a high probability, for every $i \in [n]$, $\left|T_{\bar{U}_i,i}\right| = O(\varepsilon/\sqrt{d}) = O(\varepsilon/n^2)$, since $d \geq \Omega(n^4\log(n))$. This observation motivates us to define the set $\mathcal{I} \subseteq [n]$ as follows: $i \in \mathcal{I}$ if and only if $\max\{T_{1,i}, T_{0,i}\} > \tau$ and $\min\{T_{1,i}, T_{0,i}\} < \tau$, where $\tau = \Theta(\varepsilon/n)$. We show that the expected cardinality of $\mathcal{I}$ is a lower bound on $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$. The next step of the proof is using fingerprinting lemma in Lemma 7.3 to further lower bound $|\mathcal{I}|$. We show in Lemma B.4 that we can lowerbound $|\mathcal{I}|$ using the sample-wise correlation random variables. More precisely, we show that with a high probability, $|\mathcal{I}| = \Omega\left(\left(\sum_{i=1}^n T_{U_i,i}\right)^2 / \sum_{i=1}^n T_{U_i,i}^2\right)$. Using Lemma 7.3, we show $\left(\sum_{i=1}^n T_{U_i,i}\right)^2 = \Omega(\varepsilon^2)$. Also, using Lemma B.7, we show that $\sum_{i=1}^n T_{U_i,i}^2 = O(\varepsilon^4)$. Combining these two pieces concludes the proof. For a detailed proof see Appendix C. □

6.2. Lower Bound for Improper (Unconstrained) Learners

The output of proper learners are constrained into the ball of radius one in $\mathbb{R}^d$. In this section, we prove that the lowerbound for improper (unconstrained) learners is reducible to the lowerbound for proper (constrained) learners. Consider $\mathcal{P}_{\text{cvx}}^{(d)} = (\Theta, \mathcal{Z}, f)$ described in Section 6.1.1. Using f, we define a new loss function that is supported on $\mathbb{R}^d$ as follows: for every $z \in \mathcal{Z}$, $\tilde{f} : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is given by

$$\tilde{f}(\theta, z) = \inf_{w \in \mathcal{B}_d(1)} \{f(w,z) + \|\theta - w\|\}. \tag{2}$$

Let $\mathcal{P}_{\text{cvx,improper}}^{(d)} = (\Theta, \mathcal{Z}, \tilde{f})$. From Lemma B.2, we know that $\tilde{f}(\cdot, z)$ is a 1-Lipschitz and convex function which means $\mathcal{P}_{\text{cvx,improper}}^{(d)} \in \mathcal{C}_{1,1}$.

Theorem 6.4. *Fix $\varepsilon > 0$ and let $\mathcal{P}_{\text{cvx,improper}}^{(d)}$ be as described in Section 6.2. For every $\varepsilon \leq 1$ and $\delta \leq \varepsilon$ and for every algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ that ε-learns $\mathcal{P}_{\text{cvx,improper}}^{(d)}$ with the sample complexity $N(\cdot,\cdot)$ the following holds: for every $n \geq N(\varepsilon,\delta)$, and $d \geq \Omega(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that*

$$\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) = \Omega\left(\frac{1}{\varepsilon^2}\right).$$

Proof. Let $\mathcal{A} = \{\mathcal{A}_n\}_{n\geq 1}$ be a possibly improper learning algorithm, i.e., $\mathcal{A}_n$ is not restricted to output an element of $\mathcal{B}_d(1)$. Also, let $\Pi(\mathcal{A}) = \{\Pi(\mathcal{A})_n\}_{n\geq 1}$ as a new learning algorithm that is defined as follows: for a training set $S_n \in \mathcal{Z}^n$, we have $\Pi(\mathcal{A}_n)(S_n) = \Pi(\mathcal{A}_n(S_n))$ where $\Pi(\cdot) : \mathbb{R}^d \to \mathcal{B}_d(1)$ is the orthogonal projection matrix onto $\mathcal{B}_d(1)$. Informally, $\Pi(\mathcal{A}_n)$ is based on projecting the output $\mathcal{A}_n$ to $\mathcal{B}_d(1)$. Define $\tilde{\mathrm{F}}_{\mathcal{D}}(\theta) = \mathbb{E}_{Z\sim\mathcal{D}}[\tilde{f}(\theta, Z)]$. From Lemma D.1, we know that $\hat{\theta} = \mathcal{A}_n(S_n)$ with probability one satisfies

$$\tilde{\mathrm{F}}_{\mathcal{D}}(\hat{\theta}) - \min_{\theta \in \mathcal{B}_d(1)} \tilde{\mathrm{F}}(\theta) \geq \mathrm{F}_{\mathcal{D}}(\Pi(\hat{\theta})) - \min_{\theta \in \mathcal{B}_d(1)} \mathrm{F}_{\mathcal{D}}(\theta).$$

The implication of this equation is the following: if $\mathcal{A}$ is an ε-learner for $\mathcal{P}_{\text{cvx,improper}}^{(d)}$, then, $\Pi\left(\mathcal{A}\right)$ is an ε-learner with respect to $\mathcal{P}_{\text{cvx}}^{(d)}$.

Notice that $\Pi\left(\mathcal{A}_n\right)$ is a *proper* learning algorithm. Therefore, by Theorem 4.1, we have that there exists $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that $\mathrm{CMI}_{\mathcal{D}}\left(\Pi\left(\mathcal{A}_n\right)\right) \geq \Omega\left(\frac{1}{\varepsilon^2}\right)$. Also, by Lemma D.2 (data processing inequality), $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \mathrm{CMI}_{\mathcal{D}}\left(\Pi\left(\mathcal{A}_n\right)\right)$. Ergo, for distribution $\mathcal{D}$ we also have $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega\left(\frac{1}{\varepsilon^2}\right)$. □

6.3. Matching Upper Bound

Theorem 6.5. *For every $L \in \mathbb{R}$, $R \in \mathbb{R}$, there exists a proper ε-learner with sample complexity $N(\varepsilon,\delta) = \frac{128(LR)^2}{\varepsilon^2}\log(2/\delta)$ such that the following holds: for every $0 < \delta \leq 1$, every $n \geq N(\varepsilon,\delta)$, every $(\Theta, \mathcal{Z}, f) \in \mathcal{C}_{L,R}$ and every $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ the following holds: 1) $\mathrm{F}_{\mathcal{D}}(\mathcal{A}(S_n)) - \min_{\theta\in\Theta} \mathrm{F}_{\mathcal{D}}(\theta) \leq \varepsilon$ with probability at least $1-\delta$ and 2)* $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{128(LR)^2}{\varepsilon^2}\log(2/\delta)$.

7. Characterization of CMI for the CSL SCOs

In this section, we discuss the characterization of CMI of ε-learners for CSL subclasses of SCOs. (All proofs appear in Appendix E.)

7.1. Lower Bound

7.1.1. Construction of a Hard Problem Instance

Towards proving Theorem 4.2, we develop the following construction: Let $d \in \mathbb{N}$. Let $\mathcal{Z} = \left\{\pm 1/\sqrt{d}\right\}^d$ and $\Theta = \mathbb{R}^d$. Define the loss function $f : \Theta \times \mathcal{Z} \to \mathbb{R}$ as

$f(\theta, z) = -\langle \theta, z \rangle + \frac{1}{2}\|\theta\|^2$. Let $\mathcal{P}_{\text{scvx}}^{(d)} \triangleq (\Theta, \mathcal{Z}, f)$ be the described problem instance.

7.1.2. Properties of ε-Learners for CLS

In the next lemma, we show some properties that are shared between every ε-learner of $\mathcal{P}_{\text{scvx}}^{(d)}$.

Lemma 7.1. *Fix $\varepsilon > 0$. Let $\mathcal{A}$ be an ε-learner for $\mathcal{P}_{scvx}^{(d)}$ with the sample complexity of $N(\cdot, \cdot)$ such that its output is an element of $\mathcal{B}_d(1)$. Then, for every $\delta > 0$, $n \geq N(\varepsilon, \delta)$ and every $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, with probability at least $1 - \delta$, we have $\left\|\hat{\theta} - \mu\right\|^2 \leq 2\varepsilon$, $\frac{1}{2}\|\mu\|^2 - \varepsilon \leq \left\langle \hat{\theta}, \mu \right\rangle$, and $\mathbb{E}\left[\left\langle \hat{\theta}, \mu \right\rangle\right] \geq \frac{1}{2}\|\mu\|^2 - \varepsilon - \frac{3\delta}{2}.$, where $\hat{\theta} = \mathcal{A}(S_n)$ and $\mu = \mathbb{E}_{Z \sim \mathcal{D}}[Z]$.*

Remark 7.2. For learners of $\mathcal{P}_{\text{scvx}}^{(d)}$, without loss of generality, we assume that the output of learning algorithm lies in $\mathcal{B}_d(1)$ where $\mathcal{B}_d(1)$ is the ball of radius one in $\mathbb{R}^d$. It is because, for every $\hat{\theta} \in \mathbb{R}^d$, $\mathrm{F}_{\mathcal{D}}(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} \mathrm{F}_{\mathcal{D}}(\theta) = \frac{1}{2}\left\|\hat{\theta} - \mu\right\|^2$. By Pythagorean theorem, $\left\|\Pi\left(\hat{\theta}\right) - \mu\right\|^2 \leq \left\|\hat{\theta} - \mu\right\|^2$ since $\mu \in \mathcal{B}_d(1)$ which shows that by projecting the output of any algorithm to $\mathcal{B}_d(1)$, the excess error does not increase. Notice that projection never increases CMI due to data processing inequality [CT12]. Therefore, it suffices to consider the algorithms whose output lies in $\mathcal{B}_d(1)$. ◁

The next lemma is a variant of fingerprinting lemma by Steinke [Ste16] which shows for a sufficiently accurate learner, there exists a distribution such that the correlation of the output and the training samples are bounded below by a constant.

Lemma 7.3. *Fix $\varepsilon > 0$. For every ε-learner $\mathcal{A}$ for $\mathcal{P}_{scvx}^{(d)}$ with sample complexity $N(\cdot, \cdot)$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that the following holds: for every $\delta > 0$ and $n \geq N(\varepsilon, \delta)$, let $S_n = (Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}$, $\hat{\theta} = \mathcal{A}_n(S_n)$ and $\mu = \mathbb{E}_{Z \sim \mathcal{D}}[Z]$. Then, we have*

$$\mathbb{E}\left[\sum_{i=1}^{n} \left\langle \hat{\theta} - \mu, Z_i - \mu \right\rangle\right] \geq \frac{1}{3} - 2\varepsilon - 3\delta.$$

7.1.3. CMI-Accuracy Tradeoff for CSL

Theorem (Restatement of Theorem 4.2)**.** *Let $\mathcal{P}_{scvx}^{(d)}$ be the problem instance described in Section 7.1.1. For every $\varepsilon < 1/24$ and $\delta < 1/48$ and for every ε-learner ($\mathcal{A} = \{\mathcal{A}_n\}_{n \in \mathbb{N}}$), with sample complexity $N(\cdot, \cdot)$ the following holds: for every $n \geq N(\varepsilon, \delta)$, $\delta < O(1/n^2)$, and $d \geq O(n^4 \log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that*

$$\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega\left(\frac{1}{\varepsilon}\right).$$

Proof Sketch. Let $\mathcal{P}_{\text{scvx}}^{(d)}$ be the problem instance described in Section 7.1.1. Fix an ε-learner $\mathcal{A}$ for $\mathcal{P}_{\text{scvx}}^{(d)}$, and let the data distribution be such that it satisfies Lemma 7.3. Consider the structure introduced in the definition of CMI in Definition 3.2 and define for every $i \in [n]$, $T_{0,i} = \left\langle \hat{\theta} - \mu, Z_{0,i} - \mu \right\rangle$ and $T_{1,i} = \left\langle \hat{\theta} - \mu, Z_{1,i} - \mu \right\rangle$. Let $\bar{U}_i = 1 - U_i$. An important observation is that $Z_{\bar{U}_i, i} \perp\!\!\!\perp \hat{\theta}$ given U_i. We show that $T_{\bar{U}_i, i}$ is a sub-Gaussian random variable with variance proxy of $O(1/\sqrt{d})$. Therefore, with a high probability, for every $i \in [n]$, $\left|T_{\bar{U}_i, i}\right| = O(1/\sqrt{d}) = O(1/n^2)$, since $d \geq \Omega(n^4 \log(n))$. This observation motivates us to define the set $\mathcal{I} \subseteq [n]$ as follows: $i \in \mathcal{I}$ if and only if $\max\{T_{1,i}, T_{0,i}\} > \tau$ and $\min\{T_{1,i}, T_{0,i}\} < \tau$, where $\tau = \Theta(1/n)$. We show that the expected cardinality of $\mathcal{I}$ is a lower bound on $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$. The next step of the proof is using fingerprinting lemma in Lemma 7.3 to further lower bound $|\mathcal{I}|$. Using Lemma B.4, we show that with a high probability, $|\mathcal{I}| = \Omega\left((\sum_{i=1}^n T_{U_i,i})^2 / \sum_{i=1}^n T_{U_i,i}^2\right)$. Using Lemma 7.3, we show $(\sum_{i=1}^n T_{U_i,i})^2 = \Omega(1)$. Also, using Lemma B.7, we show $\sum_{i=1}^n T_{U_i,i}^2 = O(\varepsilon)$. Combining these two pieces concludes the proof. For a detailed proof see Appendix E. □

7.2. Matching Upperbound

Theorem 7.4. *For every $L \in \mathbb{R}$, $\mu \in \mathbb{R}$, and $\varepsilon > 0$, there exists an algorithm such that the following holds: for every $(\Theta, \mathcal{Z}, f) \in \mathcal{C}_{L,\lambda}$ and for every $n \geq \frac{2L^2}{\mu\varepsilon}$, we have $\mathbb{E}[\mathrm{F}_{\mathcal{D}}(\mathcal{A}(S_n))] - \min_{\theta \in \Theta} \mathrm{F}_{\mathcal{D}}(\theta) \leq \varepsilon$, and $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{4L^2}{\mu\varepsilon}$.*

Impact Statement

This work builds upon the community's understanding of generalization error for machine learning methods. This has a positive impact on the scientific advancement of the field, and may lead to further improvements in our understanding, methodologies and applications of machine learning and AI.

Acknowledgments

The authors would like to thank Jonathan Ullman for illustrating discussions on Fingerprinting codes, Konstantina Bairaktari and Jad Silbak for helpful feedback on the drafts of this work. DMR is supported by a Canada CIFAR AI Chair and an NSERC Discovery Grant. Also, the authors would like to thank Jeffrey Negrea for his valuable suggestions during the preparation of this paper.

Disclosure of Funding

IA is supported by the Vatat Scholarship from the Israeli Council for Higher Education. MH is supported by the Khoury college distinguished postdoctoral fellowship. RL is supported by a Google fellowship, and this research has been funded, in parts, by an ERC grant (FoG - 101116258). DMR is supported by an NSERC Discovery Grant and funding through his Canada CIFAR AI Chair at the Vector Institute.

References

[Alq21] P. Alquier. *User-friendly introduction to PAC-Bayes bounds*. 2021. arXiv: 2110.11216.

[ACKL21] I. Amir, Y. Carmon, T. Koren, and R. Livni. "Never go full batch (in stochastic convex optimization)". *Advances in Neural Information Processing Systems* 34 (2021), pp. 25033–25043.

[AKL21] I. Amir, T. Koren, and R. Livni. "SGD generalizes better than GD (and regularization doesn't help)". In: *Conference on Learning Theory*. PMLR. 2021, pp. 63–92.

[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.

[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.

[BS95] D. Boneh and J. Shaw. "Collusion-secure fingerprinting for digital data". In: *Annual International Cryptology Conference*. Springer. 1995, pp. 452–465.

[BE02] O. Bousquet and A. Elisseeff. "Stability and generalization". *The Journal of Machine Learning Research* 2 (2002), pp. 499–526.

[BBFST21] G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. "When is memorization of irrelevant training data necessary for high-accuracy learning?" In: *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*. 2021, pp. 123–132.

[BBS22] G. Brown, M. Bun, and A. Smith. "Strong memory lower bounds for learning natural models". In: *Conference on Learning Theory*. PMLR. 2022, pp. 4989–5029.

[BUV14] M. Bun, J. Ullman, and S. Vadhan. "Fingerprinting codes and the price of approximate differential privacy". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014, pp. 1–10.

[CCNST+22] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. "Membership inference attacks from first principles". In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 1897–1914.

[CIJLT+22] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. *Quantifying memorization across neural language models*. 2022. arXiv: 2202.07646.

[CLEKS19] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. "The secret sharer: Evaluating and testing unintended memorization in neural networks". In: *28th USENIX Security Symposium (USENIX Security 19)*. 2019, pp. 267–284.

[CTWJH+21] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. "Extracting training data from large language models". In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2633–2650.

[CLY23] D. Carmon, R. Livni, and A. Yehudayoff. *The Sample Complexity Of ERMs In Stochastic Convex Optimization*. 2023. arXiv: 2311.05398.

[CM78] S. Cobzas and C. Mustata. "Norm preserving extension of convex Lipschitz functions". *J. Approx. theory* 24.3 (1978), pp. 236–244.

[CT12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[DFKL20] A. Dauber, M. Feder, T. Koren, and R. Livni. "Can implicit bias explain generalization? stochastic convex optimization as a case study". *Advances in Neural Information Processing Systems* 33 (2020), pp. 7743–7753.

[DSSUV15] C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. "Robust Traceability from Trace Amounts". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. 2015, pp. 650–669.

[Fel16] V. Feldman. "Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016.

[Fel20] V. Feldman. "Does learning require memorization? a short tale about a long tail". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 954–959.

[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.

[FZ20] V. Feldman and C. Zhang. "What neural networks memorize and why: Discovering the long tail via influence estimation". *Advances in Neural Information Processing Systems* 33 (2020), pp. 2881–2891.

[GSZ21] P. Grunwald, T. Steinke, and L. Zakynthinou. "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.

[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).

[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.

[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". *Advances in Neural Information Processing Systems* 33 (2020), pp. 9925–9935.

[HRTSR+23] M. Haghifam, B. Rodriguez-Galvez, R. Thobaben, M. Skoglund, D. M. Roy, and G. K. Dziugaite. "Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2023, pp. 663–706.

[HVYSI22] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani. "Reconstructing training data from trained neural networks". *Advances in Neural Information Processing Systems* 35 (2022), pp. 22911–22924.

[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).

[HD22] F. Hellström and G. Durisi. "Evaluated CMI bounds for meta learning: Tightness and expressiveness". *Advances in Neural Information Processing Systems* 35 (2022), pp. 20648–20660.

[HDGR23] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky. *Generalization bounds: Perspectives from information theory and PAC-Bayes*. 2023. arXiv: 2309.04381.

[JNGKJ19] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. *A Short Note on Concentration Inequalities for Random Vectors with Sub-Gaussian Norm*. 2019. arXiv: 1902.03736 [math.PR].

[KLSU19] G. Kamath, J. Li, V. Singhal, and J. Ullman. "Privately learning high-dimensional distributions". In: *Conference on Learning Theory*. PMLR. 2019, pp. 1853–1902.

[KLMS22] T. Koren, R. Livni, Y. Mansour, and U. Sherman. "Benign underfitting of stochastic gradient descent". *Advances in Neural Information Processing Systems* 35 (2022), pp. 19605–19617.

[LW86] N. Littlestone and M. Warmuth. "Relating data compression and learnability" (1986).

[Liv23] R. Livni. "Information Theoretic Lower Bounds for Information Theoretic Upper Bounds" (2023). arXiv: 2302.04925.

[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.

[MY16] S. Moran and A. Yehudayoff. "Sample compression schemes for VC classes". *Journal of the ACM (JACM)* 63.3 (2016), pp. 1–10.

[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.

[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.

[Ora19] F. Orabona. *A modern introduction to online learning*. 2019. arXiv: 1912.13213.

[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.

[RZ15] D. Russo and J. Zou. *How much does your data exploration overfit? Controlling bias via information usage*. 2015. arXiv: 1511.05219.

[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.

[SB14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SSSS09] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. "Stochastic Convex Optimization." In: *COLT*. Vol. 2. 4. 2009, p. 5.

[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.

[SZ20] T. Steinke and L. Zakynthinou. "Reasoning about generalization via conditional mutual information". In: *Conference on Learning Theory*. PMLR. 2020, pp. 3437–3452.

[Ste16] T. A. Steinke. "Upper and Lower Bounds for Privacy and Adaptivity in Algorithmic Data Analysis". PhD thesis. 2016.

[Tar08] G. Tardos. "Optimal probabilistic fingerprint codes". *Journal of the ACM (JACM)* 55.2 (2008), pp. 1–24.

[WM23] Z. Wang and Y. Mao. "Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[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.

[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].

[ZTL22] R. Zhou, C. Tian, and T. Liu. "Individually Conditional Individual Mutual Information Bound on Generalization Error". *IEEE Transactions on Information Theory* 68.5 (2022), pp. 3304–3316.

[Zin03] M. Zinkevich. "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th international conference on machine learning (icml-03)*. 2003, pp. 928–936.

A. Preliminaries

Notations Let $d \in \mathbb{N}$. For $x \in \mathbb{R}^d$, $\|x\|$ denotes ℓ_2 norm of x, and $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$. We denote the k-th coordinate of a d-dimensional vector x by the superscript $x^{(k)}$. For a matrix $A \in \mathbb{R}^{n \times m}$, $\|A\|_2$ is the operator norm of A. $\mathcal{B}_d(1)$ denote the ball of radius one in $\mathbb{R}^d$. For a (measurable) space $\mathcal{R}$, $\mathcal{M}_1(\mathcal{R})$ denotes the set of all probability measures on $\mathcal{R}$. Finally, let $\mathbb{1}[\cdot]$ denote the indicator function: $\mathbb{1}[p] = 1$ if predicate p is true, and $\mathbb{1}[p] = 0$ otherwise.

A.1. Background on Information Theory

Let P, Q be probability measures on a measurable space. When Q is absolutely continuous with respect to P, denoted $Q \ll P$, we write $\frac{\mathrm{d}Q}{\mathrm{d}P}$ 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 $\mathrm{KL}(Q \,\|\, P)$, equals $\int \log \frac{\mathrm{d}Q}{\mathrm{d}P} \mathrm{d}Q$ when $Q \ll P$, and is infinity otherwise. The *mutual information between X and Y* is

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

where $\otimes$ forms the product measure. The *disintegrated mutual information between X and Y given Z* is

$$I^Z(X;Y) = \mathrm{KL}(\mathbb{P}\left((X,Y)\middle|Z\right) \,\|\, \mathbb{P}\left(X\middle|Z\right) \otimes \mathbb{P}\left(Y\middle|Z\right)),$$

where $\mathbb{P}\left(Y\middle|Z\right)$ is the conditional distribution of Y given Z. Then, the conditional mutual information is

$$I(X;Y|Z) = \mathbb{E}[I^Z(X;Y)].$$

If X concentrates on a countable set V with counting measure ν, the *(Shannon) entropy of X* is $\mathrm{H}(X) = -\sum_{x \in V} \mathbb{P}(X = x) \log \mathbb{P}(X = x)$. The *disintegrated entropy of X given Y* is defined by $\mathrm{H}^Y(X) = -\sum_{x \in V} \mathbb{P}\left(X = x\middle|Y\right) \log \mathbb{P}\left(X = x\middle|Y\right)$, while the *conditional entropy of X given Y* is $\mathrm{H}(X|Y) = \mathbb{E}[\mathrm{H}^Y(X)]$.

B. Technical Lemmas

Lemma B.1 ([CT12]). *Let X and Y be discrete random variables. Then*

$$\mathrm{H}(X|Y) \le \mathrm{H}_b(\mathsf{P}_\mathsf{e}) + \mathsf{P}_\mathsf{e}\mathrm{H}(X) \le 1 + \mathsf{P}_\mathsf{e}\mathrm{H}(X),$$

where $\mathsf{P}_\mathsf{e} = \mathbb{P}(\Psi(Y) \neq X)$ for any (possibly randomized) estimator Ψ of X using Y.

Lemma B.2 (Cobzas and Mustata [CM78]). *Let $\mathcal{K}$ be a closed and convex subset of $\mathbb{R}^d$. Let $h : \mathcal{K} \to \mathbb{R}$ be a convex and L-Lipschitz function. Define $\tilde{h} : \mathbb{R}^d \to \mathbb{R}$ as*

$$\tilde{h}(x) \triangleq \inf_{y \in \mathcal{K}} \{h(y) + L\|x - y\|\}.$$

Then, we have, 1) $\tilde{h}$ is a convex and L-Lipschitz function, 2) for every $x \in \mathcal{K}$, $\tilde{h}(x) = h(x)$.

Lemma B.3. *Let X be a random variable supported on $\mathbb{R}$ with a bounded second moment. Then, for every $\theta \in \mathbb{R}$,*

$$\mathbb{P}\left(X \ge \theta\right) \ge \frac{\left(\max\{\mathbb{E}[X] - \theta, 0\}\right)^2}{\mathbb{E}[X^2]}$$

Proof. This is a non-standard variant of Paley-Zygmund inequality. With probability one,

$$\begin{aligned} X &= X\mathbb{1}\left[X < \theta\right] + X\mathbb{1}\left[X \ge \theta\right] \\ &\le \theta + X\mathbb{1}\left[X \ge \theta\right]. \end{aligned}$$

Taking an expectation and using Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}[X] \le \theta + \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{P}\left(X \ge \theta\right)} \Rightarrow \max\{\mathbb{E}[X] - \theta, 0\} \le \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{P}\left(X \ge \theta\right)},$$

which was to be shown. □

Lemma B.4. *Fix $n \in \mathbb{N}$ and $(a_1,\dots,a_n) \in \mathbb{R}^n$. Let $\sum_{i\in[n]} a_i = A_1$ and $\sum_{i\in[n]}(a_i)^2 = A_2$. Then, for every $\beta \in \mathbb{R}$,*

$$\left|\{i \in [n] \ : \ a_i \geq \beta/n\}\right| \geq \frac{\left(\max\{A_1-\beta,0\}\right)^2}{A_2}.$$

Proof. Define random variable X with the distribution Unif$(\{a_1,\dots,a_n\})$. By assumptions, $\mathbb{E}\left[X\right] = A_1/n$ and $\mathbb{E}\left[X^2\right] = A_2/n$. Notice that

$$\left|\{i \in [n] \ : \ a_i \geq \beta/n\}\right| = n\mathbb{P}\left(X \geq \beta/n\right).$$

By Lemma B.3, we have

$$\mathbb{P}\left(X \geq \beta/n\right) \geq \frac{(\max\{n\mathbb{E}[X]-\beta,0\})^2}{n^2\mathbb{E}[X^2]}.$$

Therefore,

$$\begin{aligned}\left|\{i \in [n] \ : \ a_i \geq \beta/n\}\right| &\geq \frac{(\max\{n\mathbb{E}[X]-\beta,0\})^2}{n\mathbb{E}[X^2]}\\ &= \frac{\left(\max\{A_1-\beta,0\}\right)^2}{A_2},\end{aligned}$$

as was to be shown. □

Lemma B.5. *Let $d \in \mathbb{N}$. Let $\mathcal{D} \in \mathcal{M}_1\left(\left\{\pm 1/\sqrt{d}\right\}^d\right)$ be a product distribution. Let $\mu = \mathbb{E}_{Z\sim\mathcal{D}}[Z]$. Also, let $(X_1,\dots,X_n) \sim \mathcal{D}^{\otimes n}$. Then, $\frac{1}{n}\sum_{i=1}^n (X_i-\mu)$ is a $\sqrt{1/(dn)}$ subguassian random vector. Moreover,*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right\|^2 \geq \varepsilon\right) \leq 2\exp\left(\frac{-\varepsilon n}{2}\right)$$

Proof. Let $v \in \mathbb{R}^d$ be a fixed vector and $\lambda \in \mathbb{R}$ be a constant. Then,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n \langle (X_i-\mu), v\rangle\right)\right] &= \mathbb{E}\left[\prod_{i=1}^n\prod_{k=1}^d \exp\left(\frac{\lambda}{n}\left(Z_i^{(k)}-\mu^{(k)}\right)\cdot v^{(k)}\right)\right]\\ &\leq \prod_{i=1}^n\prod_{k=1}^d \exp\left(\frac{\lambda^2 (v^{(k)})^2}{2dn^2}\right)\\ &= \exp\left(\frac{\lambda^2\left\|v\right\|^2}{2dn}\right),\end{aligned}$$

where the second step follows from Hoeffeding's Lemma. Therefore, by definition, we have the stated result. The statement regarding the concentration of the norm follows from [JNGKJ19, Lemma. 1]. □

Lemma B.6. *Fix $\beta \in [0,1]$. Let $\mu = \frac{1}{\sqrt{d}}\left(p^1,\dots,p^d\right) \in \mathbb{R}^d$ where $p = \left(p^1,\dots,p^d\right)$ is drawn from $\pi = (\text{Unif}[-\beta,\beta])^{\otimes d}$. Then,*

$$\mathbb{E}\left[\|\mu\|\right] \geq \frac{\beta}{3}.$$

Proof. We have $\mathbb{E}[(p^i)^2] = \frac{\beta^2}{3}$ for every $i \in [d]$. Notice that $\|\mu\| = \frac{1}{\sqrt{d}}\sqrt{\sum_{i=1}^d (p^i)^2}$ and for every $i \in [d]$, $(p^i)^2 \in [0,\beta^2]$ with probability one. We can write

$$\left\|p\right\|^2 = \left\|p\right\|\left\|p\right\| \leq \beta\sqrt{d}\left\|p\right\|.$$

Therefore, we have

$$\mathbb{E}[\|p\|^2] \leq \beta\sqrt{d}\mathbb{E}[\|p\|] \Rightarrow \mathbb{E}[\|p\|] \geq \frac{1}{\beta\sqrt{d}}\sum_{i=1}^d \mathbb{E}[(p^{(i)})^2] = \frac{\beta}{3}\sqrt{d}.$$

The stated result follows from $\mathbb{E}[\|\mu\|] = \frac{1}{\sqrt{d}}\mathbb{E}[\|p\|]$. □

Lemma B.7. *Let $\mathcal{Z} = \left\{\pm\frac{1}{\sqrt{d}}\right\}^d$ and $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be a product distribution and $\mu = \mathbb{E}_{Z\sim\mathcal{D}}[Z]$. Let $(Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}$ be n i.i.d. samples. Then, for every $\beta > 0$ if $d \geq \max\{32\log(2n/\beta), 32n^4\log(4n^2/\beta)\}$ with probability at least $1-\beta$ for every $y \in \mathbb{R}^d$,*

$$\sum_{i=1}^{n} \left(\langle y, Z_i - \mu\rangle\right)^2 \leq 6\left\|y\right\|^2.$$

Proof. For $i \in [n]$, define $v_i = Z_i - \mu$. Notice that $\|v_i\| = \|Z_i - \mu\| \leq 2$ since $\|Z_i\| \leq 1$ and $\|\mu\| \leq 1$. A simple calculation shows that

$$\left\|y - \sum_{i=1}^{n}\langle y, v_i\rangle v_i\right\|^2 = \|y\|^2 + \sum_{i=1}^{n}(\langle y, v_i\rangle)^2\left(\|v_i\|^2 - 2\right) + \sum_{i\neq j}\langle y, v_i\rangle\langle y, v_j\rangle\langle v_i, v_j\rangle.$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{n}(\langle y, v_i\rangle)^2\left(2 - \|v_i\|^2\right) &\leq \|y\|^2 + \sum_{i\neq j}\langle y, v_i\rangle\langle y, v_j\rangle\langle v_i, v_j\rangle \\ &\leq \|y\|^2 + |\sum_{i\neq j}\langle y, v_i\rangle\langle y, v_j\rangle\langle v_i, v_j\rangle| \\ &\leq \|y\|^2\left(1 + 4\sum_{i\neq j}|\langle v_i, v_j\rangle|\right),\end{aligned}$$

where the last step follows from Cauchy-Schwarz inequality. Consider the following events. For $i \in [n]$, define

$$\mathcal{E}_i^{(1)} = \{\|v_i\|^2 \leq 1 + \alpha\}.$$

Also, for $i, j \in [n]$ such that $i \neq j$, define the following events

$$\mathcal{E}_{i,j}^{(2)} = \{|\langle v_i, v_j\rangle| \leq \alpha/n^2\}.$$

For the first event, we have

$$\begin{aligned}1 - \mathbb{P}\left(\mathcal{E}_i^{(1)}\right) &= \mathbb{P}\left(\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2 > 1 + \alpha\right) \\ &= \mathbb{P}\left(\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2 - \mathbb{E}\left[\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2\right] > 1 + \alpha - \mathbb{E}\left[\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2\right]\right) \\ &\leq \mathbb{P}\left(\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2 - \mathbb{E}\left[\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2\right] > \alpha\right),\end{aligned}$$

where the last line follows from the fact that $\mathbb{E}\left[\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2\right] \leq 1$ since $Z_i^{(k)} \in \{\pm\frac{1}{\sqrt{d}}\}$. Then, by Hoeffding's inequality and the fact that $\left(Z_i^{(k)} - \mu^{(k)}\right)^2 \in [0, 4/d]$,

$$\mathbb{P}\left(\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2 - \mathbb{E}\left[\sum_{k=1}^{d}\left(Z_i^{(k)} - \mu^{(k)}\right)^2\right] > \alpha\right) \leq \exp\left(-\frac{d\alpha^2}{8}\right).$$

For the second type of events, since $v_i \perp\!\!\!\perp v_j$, $\|v_i\| \leq 2$, and $\|v_j\| \leq 2$ by Lemma B.8,

$$1 - \mathbb{P}\left(\mathcal{E}_{i,j}^{(2)}\right) \leq 2\exp\left(-\frac{d\alpha^2}{8n^4}\right).$$

The last step is using union bound. Set $\alpha = 1/2$ and assume $\mathbb{P}\left(\bigcup_{i\in[n]}\left(\mathcal{E}_i^{(1)}\right)^c\right) \leq n\mathbb{P}\left(\left(\mathcal{E}_1^{(1)}\right)^c\right) \leq \beta/2$. Also, $\mathbb{P}\left(\bigcup_{i\in[n]\neq j\in[n]}\left(\mathcal{E}_{i,j}^{(2)}\right)^c\right) \leq n^2\mathbb{P}\left(\left(\mathcal{E}_{1,2}^{(2)}\right)^c\right) \leq \beta/2$. Under this event,

$$\sum_{i=1}^n (\langle y, v_i\rangle)^2 (1-\alpha) \leq \|y\|^2 (1+4\alpha) \Rightarrow \sum_{i=1}^n (\langle y, v_i\rangle)^2 \leq 6\|y\|^2,$$

which was to be shown. □

Lemma B.8. *Let $\mathcal{Z} = \{\pm\frac{1}{\sqrt{d}}\}^d$ and $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be a product measure. Define $\mu = \mathbb{E}_{Z\sim\mathcal{D}}[Z]$. Then, for every fixed $y \in \mathbb{R}^d$ and $n \in \mathbb{N}$,*

$$\mathbb{P}_{(Z_1,\dots,Z_n)\sim\mathcal{D}^{\otimes n}}\left(\max_{i\in[n]}\{\langle y, Z_i - \mu\rangle\} \geq \alpha\right) \leq n\exp\left(-\frac{\alpha^2 d}{2\|y\|^2}\right).$$

Proof. By union bound, $\mathbb{P}\left(\max_{i\in[n]}\{\langle y, Z_i-\mu\rangle\} \geq \alpha\right) \leq n\mathbb{P}_{Z\sim\mathcal{D}}\left(\langle y, Z-\mu\rangle \geq \alpha\right)$. Let $\lambda > 0$ and consider

$$\begin{aligned}
\mathbb{E}[\exp(\lambda\langle y, Z-\mu\rangle)] &= \mathbb{E}\left[\exp\left(\lambda\sum_{k=1}^d y^{(k)}\left(Z^{(k)} - \mu^{(k)}\right)\right)\right] \\
&= \prod_{k=1}^d \mathbb{E}\left[\exp\left(\lambda y^{(k)}\left(Z^{(k)} - \mu^{(k)}\right)\right)\right] \\
&\leq \prod_{k=1}^d \exp\left(\lambda^2 (y^{(k)})^2 \frac{1}{2d}\right) \quad \text{(Hoeffeding's lemma since } Z^{(k)} \in \{\pm 1/\sqrt{d}\}) \\
&= \exp\left(\lambda^2\|y\|^2\frac{1}{2d}\right).
\end{aligned}$$

Then, using standard arguments, the stated claim can be proved. □

Lemma B.9. *[SB14, Lemma B.1] Let X ba a non-negative random variable supported on $\mathbb{R}$ and $\mathbb{P}(X \leq a) = 1$. Then,*

$$\mathbb{P}(X > \beta) \geq \frac{\mathbb{E}[X] - \beta}{a - \beta}.$$

C. Proofs for Characterization of CMI of the CLB Subclass

C.1. Proof of Lemma 6.2

Notice that $\mathrm{F}_{\mathcal{D}}(\theta) = -\langle\theta, \mu\rangle$ and $\min_{\theta\in\Theta}\mathrm{F}_{\mathcal{D}}(\theta) = -\|\mu\|$, where the minimum is achieved by setting $\theta^\star = \frac{\mu}{\|\mu\|}$. Therefore, by the excess risk guarantee, with probability at least $1-\delta$,

$$\mathrm{F}_{\mathcal{D}}(\hat{\theta}) + \|\mu\| \leq \varepsilon \Rightarrow \|\mu\| - \varepsilon \leq \left\langle\hat{\theta}, \mu\right\rangle.$$

Notice that $\left\langle\hat{\theta}, \mu\right\rangle \geq -1$, $\|\mu\| \leq 1$, and $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{E}\left[\left\langle\hat{\theta}, \mu\right\rangle\right] &\geq (\|\mu\| - \varepsilon)\,\mathbb{P}\left(\left\langle\hat{\theta}, \mu\right\rangle \geq (\|\mu\| - \varepsilon)\right) - \mathbb{P}\left(\left\langle\hat{\theta}, \mu\right\rangle < (\|\mu\| - \varepsilon)\right) \\
&= (\|\mu\| - \varepsilon)\left(1 - \mathbb{P}\left(\left\langle\hat{\theta}, \mu\right\rangle < (\|\mu\| - \varepsilon)\right)\right) - \mathbb{P}\left(\left\langle\hat{\theta}, \mu\right\rangle < (\|\mu\| - \varepsilon)\right) \\
&= (\|\mu\| - \varepsilon) - \mathbb{P}\left(\left\langle\hat{\theta}, \mu\right\rangle < (\|\mu\| - \varepsilon)\right)(\|\mu\| - \varepsilon + 1) \\
&\geq (\|\mu\| - \varepsilon) - 2\delta,
\end{aligned}$$

where the last step follows because $\|\mu\| - \varepsilon + 1 \leq 2$ and $\mathbb{P}\left(\left\langle\hat{\theta}, \mu\right\rangle < (\|\mu\| - \varepsilon)\right) \leq \delta$ by the first part of the lemma.

C.2. Proof of Lemma 6.3

The proof is based on defining a family of data distribution, and a prior over the family. Then, we show that in-expectation over the prior, the stated claim holds. Thus, there exists a distribution with the desired property.

The data distribution is parameterized by a vector $p = \left(p^{(1)}, \dots, p^{(d)}\right) \in [-1,1]^d$ where for every $z = (z^{(1)}, \dots, z^{(d)}) \in \{\pm \frac{1}{\sqrt{d}}\}^d$,

$$\mathcal{D}_p(z = (z^{(1)}, \dots, z^{(d)})) = \prod_{k=1}^{d} \left(\frac{1 + \sqrt{d} z^{(k)} p^{(k)}}{2} \right).$$

Let $\mu_p = \mathbb{E}_{X \sim \mathcal{D}_p}[X]$ where $\mu_p^{(k)} = p^{(k)}/\sqrt{d}$ for $k \in [d]$.

Then we define a *prior* distribution $\pi \in \mathcal{M}_1([-1,1]^d)$ over p denoted by π and is given by

$$\pi = \text{Unif}([-12\varepsilon, 12\varepsilon])^{\otimes d}.$$

Let $S_n = (Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$, and $\hat{\theta} = \mathcal{A}_n(S_n)$. By the same proof as presented in [KLSU19] (see Equation 16 therein), we have that

$$\mathbb{E}_{p\sim\pi} \mathbb{E}_{S_n \sim \mathcal{D}_p^{\otimes n}} \left[\sum_{i=1}^{n} \sum_{k=1}^{d} \left(\frac{144\varepsilon^2 - d(\mu_p^{(k)})^2}{1 - d(\mu_p^{(k)})^2} \right) \left(\hat{\theta}^{(k)}\right) \left(Z_i^{(k)} - \mu_p^{(k)}\right) \right] = 2\mathbb{E}_{p\sim\pi} \left[\left\langle \mathbb{E}_{S_n \sim \mathcal{D}_p^{\otimes n}}[\hat{\theta}], \mu_p \right\rangle \right]. \tag{3}$$

By Lemma 6.2, we know that for every $p \in [-1,1]^d$

$$\left\langle \mathbb{E}_{S_n \sim \mathcal{D}_p^{\otimes n}}[\hat{\theta}], \mu_p \right\rangle \geq \|\mu_p\| - \varepsilon - 2\delta. \tag{4}$$

Also, by Lemma B.6, we have

$$\mathbb{E}_{p\sim\pi} \left[\|\mu_p\|\right] \geq 4\varepsilon. \tag{5}$$

Therefore, by Equations (3) to (5), we have

$$\mathbb{E}_{p\sim\pi} \mathbb{E}_{S_n \sim \mathcal{D}_p^{\otimes n}} \left[\sum_{i=1}^{n} \sum_{k=1}^{d} \left(\frac{144\varepsilon^2 - d(\mu_p^{(k)})^2}{1 - d(\mu_p^{(k)})^2} \right) \left(\hat{\theta}^{(k)}\right) \left(Z_i^{(k)} - \mu_p^{(k)}\right) \right] \geq 2\varepsilon - 4\delta,$$

which was to be shown.

C.3. Proof of Theorem 4.1

Fix a learning algorithm $\mathcal{A}$ and let $\mathcal{D}$ be a distribution satisfies Lemma 6.3. Also, consider the structure used in the definition of CMI in Definition 3.2 and let $\tilde{\boldsymbol{Z}} = \{Z_{j,i}\}_{j\in\{0,1\}, i\in[n]} \sim \mathcal{D}^{\otimes(2\times n)}$. For every $j \in \{0,1\}$ and $i \in [n]$, define $v_{j,i} = (v_{j,i}^{(1)}, \dots, v_{j,i}^{(d)}) \in \mathbb{R}^d$ as follows. For every $k \in [d]$, let

$$v_{j,i}^{(k)} \triangleq \frac{144\varepsilon^2 - d(\mu^{(k)})^2}{1 - d(\mu^{(k)})^2} \left(Z_{j,i}^{(k)} - \mu^{(k)}\right).$$

In the first step, we make the following observations. From the construction in Lemma 6.3, we know that $\mu^{(k)} \in [-12\varepsilon/\sqrt{d}, 12\varepsilon/\sqrt{d}]$. Simple calculations show,

$$\frac{144\varepsilon^2 - d(\mu^{(k)})^2}{1 - d(\mu^{(k)})^2} \leq 144\varepsilon^2. \tag{6}$$

Let $\beta \triangleq \varepsilon$ be a constant. Define the following set

$$\mathcal{I} = \left\{ (i,j) \in [n] \times \{0,1\} \middle| \left\langle \hat{\theta}, v_{j,i} \right\rangle \geq \beta/n \text{ and } \left\langle \hat{\theta}, v_{1-j,i} \right\rangle < \beta/n \right\}.$$

Intuitively, $\mathcal{I}$ includes the subset of columns of supersample such that one of the samples has a *large* correlation with the output of the algorithm and the other one has *small* correlation with the output of the algorithm. Also, define the following event

$$\mathcal{G} = \left\{\forall i \in [n] : \left\langle \hat{\theta}, v_{\bar{U}_i,i} \right\rangle < \beta/n \right\},$$

where $\bar{U}_i = 1 - U_i$. Intuitively, under the event $\mathcal{G}$ the correlation of the output and the *heldout samples* are insignificant.

We can write

$$\begin{aligned} \mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= \mathrm{H}(U|\tilde{\boldsymbol{Z}}) - \mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}) \\ &= \mathrm{H}(U) - \mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}) \\ &= n - \mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}). \end{aligned}$$

Notice that $\mathcal{I}$ is a $(\hat{\theta}, \tilde{\boldsymbol{Z}})$-measurable random variable, thus, $\mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}) = \mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I})$. Define $\mathcal{I}^{(1)}$ as follows: $i \in \mathcal{I}^{(1)}$ iff $\exists j \in \{0,1\}$ such that $(i,j) \in \mathcal{I}$. Using this notation, we can write

$$\begin{aligned} \mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}) &= \mathrm{H}(U_{\mathcal{I}^{(1)}}, U_{(\mathcal{I}^{(1)})^c}|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}) \\ &\leq \mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}) + \mathrm{H}(U_{(\mathcal{I}^{(1)})^c}|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}), \end{aligned} \tag{7}$$

where the last step follows from sub-additivity of Entropy. The second term in Equation (7) can be bounded by

$$\mathrm{H}(U_{(\mathcal{I}^{(1)})^c}|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}) \leq \mathbb{E}\left[(n - |\mathcal{I}|)\right].$$

Define the random variable $\hat{U} \in \{0,1\}^n$ as follows: for every $(i,j) \in \mathcal{I}$, let $\hat{U}_i = j$. For the remaining coordinates set $\hat{U}_i = 0$. Notice that $\hat{U}$ is a $\mathcal{I}$-measurable random variable. Therefore, $\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}) = \mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}, \hat{U})$. Then, we invoke Fano's inequality from Lemma B.1 to write

$$\begin{aligned} \mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}, \hat{U}) &\leq \mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\hat{U}) \\ &\leq 1 + \mathrm{H}(U_{\mathcal{I}^c})\mathbb{P}\left(\{\exists (i,j) \in \mathcal{I} : U_i \neq j\}\right) \\ &\leq 1 + n\mathbb{P}\left(\{\exists (i,j) \in \mathcal{I} : U_i \neq j\}\right). \end{aligned}$$

We claim that $\mathbb{P}\left(\{\exists (i,j) \in \mathcal{I} : U_i \neq j\}\right) \leq \mathbb{P}\left(\mathcal{G}^c\right)$. The proof is as follows: If there exists $(i,j) \in \mathcal{I}$ such that $U_i \neq j$, then, we have

$$\left\langle \hat{\theta}, v_{\bar{U}_i,i} \right\rangle \geq \beta/n,$$

by the definition of $\mathcal{I}$. Therefore, we conclude $\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}}, \hat{\theta}, \mathcal{I}) \leq 1 + n\mathbb{P}\left(\mathcal{G}^c\right)$. The conditional entropy can be upper-bounded by

$$\mathrm{H}(U\big|\tilde{\boldsymbol{Z}}, \hat{\theta}) \leq n - \mathbb{E}\left[|\mathcal{I}|\right] + 1 + n\mathbb{P}\left(\mathcal{G}^c\right).$$

By the definition of mutual information, we can lower bound $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ as follows

$$\begin{aligned} \mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= n - \mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}) \\ &\geq \mathbb{E}\left[|\mathcal{I}|\right] - 1 - n\mathbb{P}\left(\mathcal{G}^c\right). \end{aligned} \tag{8}$$

In the next part of the proof we will show that $\mathbb{P}\left(\mathcal{G}^c\right) = O(1/n^2)$.

In the next step of the proof, we provide a lowerbound on $|\mathcal{I}|$. Under the event $\mathcal{G}$, using Lemma B.4 we can lower bound $|\mathcal{I}|$ as follows

$$\begin{aligned} \mathbb{E}\left[|\mathcal{I}|\right] &\geq \mathbb{E}\left[|\mathcal{I}|\mathbb{1}\left[\mathcal{G}\right]\right] \\ &\geq \mathbb{E}\left[\left|\left\{i \in [n] : \left\langle \hat{\theta}, v_{U_i,i} \right\rangle \geq \frac{\beta}{n}\right\}\right| \mathbb{1}\left[\mathcal{G}\right]\right] \\ &\geq \mathbb{E}\left[\frac{\left(\max\left\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\right\}\right)^2}{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle^2} \mathbb{1}\left[\mathcal{G}\right]\right]. \end{aligned}$$

Define the following event

$$\mathcal{E} \triangleq \mathcal{G} \cap \left\{ \sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle^2 \leq 6\varepsilon^4 \right\}.$$

Since $\mathcal{E} \subseteq \mathcal{G}$, we have

$$\begin{aligned}\mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle^2} \mathbb{1}\left[\mathcal{G}\right] \right] &\geq \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle^2} \mathbb{1}\left[\mathcal{E}\right] \right] \\ &\geq \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{6\varepsilon^4} \mathbb{1}\left[\mathcal{E}\right] \right],\end{aligned}$$

where the last step follows because under the event $\mathcal{E}$, $\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle^2 \leq 6\varepsilon^4$. Then, consider

$$\begin{aligned}\mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{6\varepsilon^4} \mathbb{1}\left[\mathcal{E}\right] \right] &= \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{3\varepsilon^4} \right] \\ &\quad - \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{6\varepsilon^4} \mathbb{1}\left[\mathcal{E}^c\right] \right].\end{aligned} \tag{9}$$

The first term in Equation (9) can be lower bounded as

$$\begin{aligned}\mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{6\varepsilon^4} \right] &\geq \frac{\left(\max\{\mathbb{E}\left[\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle\right] - \beta, 0\}\right)^2}{6\varepsilon^4} \\ &\geq \frac{\left(\max\{6\varepsilon - 4\delta - \beta, 0\}\right)^2}{6\varepsilon^4}.\end{aligned} \tag{10}$$

where the first step follows from convexity of $h_1(x) = x^2$ and $h_2(x) = \max\{x, 0\}$ and applying Jensen's inequality. The second step follows from Lemma 6.3. Since $\delta < \varepsilon$ and $\beta = \varepsilon$,

$$\mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{6\varepsilon^4} \right] \geq \frac{1}{6\varepsilon^2}.$$

The second term in Equation (9) can be upperbounded by

$$2\mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\}\right)^2}{3\varepsilon^4} \mathbb{1}\left[\mathcal{E}^c\right] \right] \leq \frac{16\varepsilon^4 n^2 + 4\beta^2}{3\varepsilon^4} \cdot \mathbb{P}\left(\mathcal{E}^c\right),$$

where the last step follows from

$$\begin{aligned}\left(\max\left\{\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle - \beta, 0\right\}\right)^2 &\leq 2\left\|\hat{\theta}\right\|^2 \left\|\sum_{i\in[n]} v_{U_i,i}\right\|^2 + 2\beta^2 \\ &\leq 8\varepsilon^4 n^2 + 2\beta^2 \\ &= 8\varepsilon^4 n^2 + 2\varepsilon^2.\end{aligned}$$

Here, we use $\|v_{U_i,i}\| \leq 2\varepsilon^2$ due to Equation (6). Therefore, we need to show that for sufficiently small γ, $\mathbb{P}\left(\mathcal{E}^c\right) \leq \frac{\gamma}{n^2}$. We can use union bound to write $\mathbb{P}\left(\mathcal{E}^c\right) \leq \mathbb{P}\left(\mathcal{G}^c\right) + \mathbb{P}\left(\sum_{i\in[n]} \left\langle \hat{\theta}, v_{U_i,i} \right\rangle > 3/2\varepsilon^4\right)$. Notice that

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}^c\right) &= \mathbb{P}\left(\max_{i\in[n]} \left\{\left\langle \hat{\theta}, v_{\bar{U}_i,i} \right\rangle\right\} \geq \beta/n\right) \\ &= \mathbb{E}\left[\mathbb{P}\left(\max_{i\in[n]} \left\{\left\langle \hat{\theta}, v_{\bar{U}_i,i} \right\rangle\right\} \geq \beta/n \middle| U, \hat{\theta}\right)\right].\end{aligned}$$

Note that conditioned on U, $\hat{\theta}$ and $v_{\bar{U}_i,i}$ are independent by the construction of CMI in Definition 3.2. This observation lets us use Lemma B.8 to upperbound the probability inside the expectation. Since $\left\|\hat{\theta}\right\| \leq 1$, if $d \geq O(n^2\log(n^3))$, we have

$$\mathbb{P}\left(\mathcal{G}^c\right) \leq O\left(\frac{1}{n^2}\right). \tag{11}$$

Let us define diagonal matrix $A \in \mathbb{R}^{d\times d}$ as

$$A = \operatorname{diag}\left[\left\{\frac{144\varepsilon^2 - d(\mu^{(k)})^2}{1 - d(\mu^{(k)})^2}\right\}_{k=1}^{d}\right].$$

By Equation (6) $\|A\|_2 \leq \varepsilon^2$ due to Equation (6). By the fact that A is a symmetric matrix, we can write

$$\sum_{i\in[n]} \left\langle \hat{\theta}, A\left(Z_i - \mu\right)\right\rangle = \sum_{i\in[n]} \left\langle A\hat{\theta}, \left(Z_i - \mu\right)\right\rangle.$$

Since $\left\|A\hat{\theta}\right\|^2 \leq \|A\|^2 \left\|\hat{\theta}\right\|^2 \leq 144^2\varepsilon^4$, we can write

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} \left\langle \hat{\theta}, v_{U_i,i}\right\rangle^2 \geq 6\varepsilon^2\right) &= \mathbb{P}\left(\sum_{i=1}^{n} \left\langle A\hat{\theta}, Z_{U_i,i} - \mu\right\rangle^2 \geq 6/144^2\varepsilon^2\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{n} \left\langle A\hat{\theta}, Z_{U_i,i} - \mu\right\rangle^2 \geq 6/144^2 \left\|A\hat{\theta}\right\|^2\right) \\ &\leq \mathbb{E}\left[\mathbb{P}\left(\sum_{i=1}^{n} \left\langle A\hat{\theta}, Z_{U_i,i} - \mu\right\rangle^2 \geq 6 \left\|A\hat{\theta}\right\|^2 \middle| U\right)\right].\end{aligned}$$

Using this representation, we can use Lemma B.7 to conclude that if $d > O\left(n^4\log(n^5)\right)$

$$\mathbb{P}\left(\sum_{i=1}^{n} \left\langle A\hat{\theta}, Z_{U_i,i} - \mu\right\rangle^2 \geq \left\|A\hat{\theta}\right\|^2 \middle| U\right) \leq O\left(\frac{1}{n^2}\right). \tag{12}$$

To conclude the proof, Equation (11) and Equation (12) show

$$\mathbb{P}\left(\mathcal{E}^c\right) \leq O\left(\frac{1}{n^2}\right).$$

C.4. Proof of Theorem 6.5

Given that the Euclidean radius of Θ is bounded by R, we will presume that the loss function lies within $[-LR, LR]$. Let $0 < m \leq n$ and $\eta > 0$ be constants which are determined later. The algorithm $\mathcal{A}_n$ is based on early-stopped online gradient descent. More precisely, let the training set $S_n = (Z_1, \dots, Z_n)$ and $\theta_1 = 0$. For $t \in [m]$, let

$$\theta_{t+1} = \Pi_\Theta\left(\theta_t - \eta\partial f(\theta_t, Z_t)\right),$$

where $\partial f(\theta_t, Z_t)$ denotes the sub-gradient of $\partial f(\cdot, Z_t)$ at θ_t. Then, the output of the algorithm will be $\mathcal{A}_n(S_n) = \frac{1}{m}\sum_{t=1}^{m}\theta_t$.

By the standard result on the regret analysis of the online gradient descent and the online-to-batch conversion in [Zin03; SSSS09; Ora19], we have with probability at least $1-\delta$,

$$\mathrm{F}_{\mathcal{D}}(\mathcal{A}_n(S_n)) - \min_{\theta\in\Theta}\mathrm{F}_{\mathcal{D}}(\theta) \le \frac{R^2}{2m\eta} + \frac{\eta}{2}L^2 + 2LR\sqrt{\frac{8\log(2/\delta)}{m}}$$

By setting $m = 128\frac{(LR)^2}{\varepsilon^2}\log(2/\delta)$ and $\eta = \frac{R}{L}\frac{1}{\sqrt{m}}$, $\mathcal{A}_n$ achieves ε excess risk of ε with probability at least $1-\delta$. Next, we provide the analysis of CMI of $\mathcal{A}_n$. Then, using chain rule for mutual information, we have

$$\begin{aligned}\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= I(\mathcal{A}_n(S_n); U|\tilde{\boldsymbol{Z}})\\ &= I(\mathcal{A}_n(S_n); U_1,\dots,U_n|\tilde{\boldsymbol{Z}})\\ &= I(\mathcal{A}_n(S_n); U_1,\dots,U_m|\tilde{\boldsymbol{Z}})\\ &+ I(\mathcal{A}_n(S_n); U_{m+1},\dots,U_n|\tilde{\boldsymbol{Z}}, U_1,\dots,U_m).\end{aligned}$$

Since $\mathcal{A}_n(S_n)$ depends only on the first m examples in the training set, $I(\mathcal{A}_n(S_n); U_{m+1},\dots,U_n|\tilde{\boldsymbol{Z}}, U_1,\dots,U_m) = 0$. Therefore,

$$\begin{aligned}\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= I(\mathcal{A}_n(S); U_1,\dots,U_m|\tilde{\boldsymbol{Z}})\\ &\le \mathrm{H}(U_1,\dots,U_m|\tilde{\boldsymbol{Z}})\\ &= \mathrm{H}(U_1,\dots,U_m)\\ &\le m.\end{aligned} \tag{13}$$

Therefore its CMI is less than m as was to be shown.

C.5. Corollaries of Proof of Theorem 4.1

Corollary C.1. *Let $\mathcal{P}_{cvx}^{(d)}$ be the problem instance described in Section 6.1.1. Fix $\varepsilon < 1$. For every $\delta \le \varepsilon$ and for every algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ that ε-learns $\mathcal{P}_{cvx}^{(d)}$ with the sample complexity $N(\cdot,\cdot)$ the following holds: for every $n \ge N(\varepsilon,\delta)$, and $d \ge \Omega(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that*

$$\mathbb{E}\left[\left|\left\{i\in[n] : \left\langle\hat{\theta}, A\left(Z_i-\mu\right)\right\rangle\right\} \ge \frac{\varepsilon}{n}\right|\right] = \Omega\left(\frac{1}{\varepsilon^2}\right),$$

where $S_n = (Z_1,\dots,Z_n) \sim \mathcal{D}^{\otimes n}$, $\hat{\theta} = \mathcal{A}(S_n)$, and $\mu = \mathbb{E}_{Z\sim\mathcal{D}}[Z]$, and

$$A = \operatorname{diag}\left[\left\{\frac{\varepsilon^2-(\mu^{(k)})^2}{1-(\mu^{(k)})^2}\right\}_{k=1}^{d}\right].$$

Corollary C.2. *Fix $\varepsilon > 0$. Consider the structure introduced in the definition of CMI in Definition 3.2. Then, define the random variable*

$$\mathcal{I} = \left\{(i,j)\in[n]\times\{0,1\}\middle|\ \langle\mathcal{A}_n(S_n), A\left(Z_{j,i}-\mu\right)\rangle \ge \varepsilon/n \textit{ and } \langle\mathcal{A}_n(S_n), A\left(Z_{1-j,i}-\mu\right)\rangle < \varepsilon/n\right\},$$

where

$$A = \operatorname{diag}\left[\left\{\frac{\varepsilon^2-(\mu^{(k)})^2}{1-(\mu^{(k)})^2}\right\}_{k=1}^{d}\right].$$

Let $\mathcal{P}_{cvx}^{(d)}$ be the problem instance described in Section 6.1.1. Fix $\varepsilon < 1$. For every $\delta \le \varepsilon$ and for every algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{N}}$ that ε-learns $\mathcal{P}_{cvx}^{(d)}$ with the sample complexity $N(\cdot,\cdot)$ the following holds: for every $n \ge N(\varepsilon,\delta)$, and $d \ge \Omega(n^4\log(n))$, there exists a data distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ such that

$$\mathbb{E}\left[|\mathcal{I}|\right] = \Omega\left(\frac{1}{\varepsilon}\right).$$

D. Auxiliary Lemma for Improper Learning of the CLB Subclass

Lemma D.1. *Let $\mathcal{B}_d(1)$ denote the ball of radius one in $\mathbb{R}^d$. Let $f : \mathcal{B}_d(1) \times \mathcal{Z} \to \mathbb{R}$ be a convex and 1-Lipschitz loss function defined over $\mathcal{B}_d(1)$. Then, there exists a convex and 1-Lipschitz $\tilde{f} : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ such that for every $\hat{\theta} \in \mathbb{R}^d$ and every $\mathcal{D}$, we have*

$$\mathbb{E}_{Z\sim\mathcal{D}}\left[\tilde{f}(\hat{\theta}, Z)\right] - \min_{\theta\in\mathcal{B}_d(1)} \mathbb{E}_{Z\sim\mathcal{D}}\left[\tilde{f}(\theta, Z)\right] \geq \mathbb{E}_{Z\sim\mathcal{D}}\left[f(\Pi\left(\hat{\theta}\right), Z)\right] - \min_{\theta\in\mathcal{B}_d(1)} \mathbb{E}_{Z\sim\mathcal{D}}\left[f(\theta, Z)\right],$$

where $\Pi(\cdot) : \mathbb{R}^d \to \mathcal{B}_d(1)$ is the orthogonal projection operator on $\mathcal{B}_d(1)$.

Proof. Let $f : \mathcal{B}_d(1) \times \mathcal{Z} \to \mathbb{R}$ be a convex and 1-Lipschitz loss function. For every $z \in \mathcal{Z}$, define

$$\tilde{f}(\theta, z) = \inf_{w\in\mathcal{B}_d(1)} \{f(w, z) + \|\theta - w\|\}.$$

By Lemma B.2, we know that for every $z \in \mathcal{Z}$, $\tilde{f}(\cdot, z)$ is convex and $1-$Lipschitz. Our first claim is that

$$\min_{\theta\in\mathcal{B}_d(1)} \mathbb{E}_{Z\sim\mathcal{D}}\left[\tilde{f}(\theta, Z)\right] = \min_{\theta\in\mathcal{B}_d(1)} \mathbb{E}_{Z\sim\mathcal{D}}\left[f(\theta, Z)\right].$$

It follows from the fact that for every $\theta \in \mathcal{B}_d(1)$ and every $z \in \mathcal{Z}$, $\tilde{f}(\theta, z) = f(\theta, z)$ by Lemma B.2. Let $\Pi : \mathbb{R}^d \to \mathcal{B}_d(1)$ be the projection operator. Our second claim is that for every $\theta \in \mathbb{R}^d$, we have

$$\mathbb{E}_{Z\sim\mathcal{D}}\left[\tilde{f}(\Pi\left(\theta\right), Z)\right] \leq \mathbb{E}_{Z\sim\mathcal{D}}\left[\tilde{f}(\theta, Z)\right]. \tag{14}$$

The proof is as follows. For every $z \in \mathcal{Z}$, we can write

$$\begin{aligned}\tilde{f}\left(\theta, z\right) &= \inf_{w\in\mathcal{B}_d(1)} \{f(w, z) + \|\theta - w\|\}\\ &\geq \inf_{w\in\mathcal{B}_d(1)} \{f(w, z) + \|\Pi\left(\theta\right) - w\|\},\end{aligned}$$

where the last step follows from

$$\|\theta - w\| \geq \|\Pi(\theta) - \Pi(w)\| = \|\Pi(\theta) - w\|$$

where the first step is by contraction property of the projection and the second step is due to $\Pi(w) = w$ since $w \in \mathcal{B}_d(1)$. Then, notice that

$$\begin{aligned}\inf_{w\in\mathcal{B}_d(1)} \{f(w, z) + \|\Pi\left(\theta\right) - w\|\} &= \tilde{f}(\Pi\left(\theta\right), z)\\ &= f(\Pi\left(\theta\right), z).\end{aligned}$$

The last step follows from $\Pi(\theta) \in \mathcal{B}_d(1)$ and by Lemma B.2, $\tilde{f}(., z)$ and $f(., z)$ agree on $\mathcal{B}_d(1)$.

Combining these two claims we obtain, for every $\hat{\theta} \in \mathbb{R}^d$, we have

$$\mathbb{E}_{Z\sim\mathcal{D}}\left[\tilde{f}(\hat{\theta}, Z)\right] - \min_{\theta\in\mathcal{B}_d(1)} \mathbb{E}_{Z\sim\mathcal{D}}\left[\tilde{f}(\theta, Z)\right] \geq \mathbb{E}_{Z\sim\mathcal{D}}\left[f(\Pi\left(\hat{\theta}\right), Z)\right] - \min_{\theta\in\mathcal{B}_d(1)} \mathbb{E}_{Z\sim\mathcal{D}}\left[f(\theta, Z)\right],$$

as was to be shown. □

Lemma D.2. *Let $\mathcal{A}_n$ be a learning algorithm. Define $\Pi(\mathcal{A}_n)$ as a learning algorithm that obtains by projecting the output of $\mathcal{A}_n$ into $\mathcal{B}_d(1)$. Then,*

$$\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \mathrm{CMI}_{\mathcal{D}}(\Pi(\mathcal{A}_n))$$

Proof. This result is a direct corollary of the data processing inequality [CT12]. □

E. Proofs for Characterization of CMI of the CSL Subclass

E.1. Proof of Lemma 7.1

For every $\theta \in \mathbb{R}^d$, we have $\mathrm{F}_{\mathcal{D}}(\theta) = -\langle \theta, \mu \rangle + \frac{1}{2}\left\|\theta\right\|^2$, and $\min\limits_{\theta \in \Theta} \mathrm{F}_{\mathcal{D}}(\theta) = \frac{-1}{2}\left\|\mu\right\|^2$ where the minimum is achieved by setting $\theta^\star = \mu$. Therefore, simple calculation shows that

$$\begin{aligned}
\mathrm{F}_{\mathcal{D}}(\theta) - \mathrm{F}_{\mathcal{D}}(\theta^\star) &= \frac{1}{2}\left\|\theta - \mu\right\|^2 \\
&= \frac{1}{2}\left\|\theta\right\|^2 - \langle \theta, \mu \rangle + \frac{1}{2}\left\|\mu\right\|^2 \\
&\geq \frac{1}{2}\left\|\mu\right\|^2 - \langle \theta, \mu \rangle .
\end{aligned}$$

Thus, if $\hat{\theta}$ achieves excess error ε with probability at least $1-\delta$, we have $\frac{1}{2}\left\|\mu\right\|^2 - \left\langle \hat{\theta}, \mu \right\rangle \leq \varepsilon$ and $\left\|\theta - \mu\right\|^2 \leq 2\varepsilon$.

For the in-expectation result, notice that without loss of generality, we can assume that $\hat{\theta} \in \mathcal{B}_d(1)$.

E.2. Proof of Lemma 7.3

The proof is based on defining a family of data distribution, and a prior over the family. Then, we show that in-expectation over the prior, the stated claim holds.

The data distribution is parameterized by a vector $p = \left(p^{(1)}, \dots, p^{(d)}\right) \in [-1,1]^d$ where for every $z = (z^{(1)}, \dots, z^{(d)}) \in \{\pm \frac{1}{\sqrt{d}}\}^d$,

$$\mathcal{D}_p(z = (z^{(1)}, \dots, z^{(d)})) = \prod_{k=1}^{d} \left(\frac{1 + \sqrt{d} z^{(k)} p^{(k)}}{2} \right).$$

Let $\mu_p = \mathbb{E}_{Z \sim \mathcal{D}_p}[Z]$ where $\mu_p^{(k)} = p^{(k)}/\sqrt{d}$. We define a *prior* distribution $\pi \in \mathcal{M}_1([-1,1]^d)$ over p denoted by π and is given by

$$\pi = \mathrm{Unif}([-1,1])^{\otimes d}.$$

Let $S_n = (Z_1, \dots, Z_n) \sim \mathcal{D}_p^{\otimes n}$, and $\hat{\theta} = \mathcal{A}_n(S_n)$. From Lemmas 4.3.7 and 4.3.8 of [Ste16], we have the following result known as finerprinting lemma:

$$\begin{aligned}
&\mathbb{E}_{p \sim \pi} \mathbb{E}_{S_n \sim \mathcal{D}^{\otimes n}} \left[\left\langle \hat{\theta}, \sum_{i \in [n]} (Z_i - \mu_p) \right\rangle \right] \\
&= 2\mathbb{E}_{p \sim \pi} \left[\left\langle \mathbb{E}_{S_n \sim \mathcal{D}^{\otimes n}}[\hat{\theta}], \mu_p \right\rangle \right].
\end{aligned}$$

By Lemma 7.1, for every p

$$\left\langle \mathbb{E}_{S_n \sim \mathcal{D}_p^{\otimes n}}[\hat{\theta}], \mu_p \right\rangle \geq \frac{\left\|\mu_p\right\|^2}{2} - \varepsilon - \frac{3}{2}\delta.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{p \sim \pi} \left\langle \mathbb{E}_{S_n \sim \mathcal{D}_p^{\otimes n}}[\hat{\theta}], \mu_p \right\rangle &\geq \mathbb{E}_{p \sim \pi} \left[\frac{\left\|\mu_p\right\|^2}{2} \right] - \varepsilon - \frac{3}{2}\delta \\
&= \frac{1}{6} - \varepsilon - \frac{3}{2}\delta,
\end{aligned}$$

where the last step follows from

$$\mathbb{E}_{p \sim \pi}\left[\left\|\mu_p\right\|^2 \right] = \sum_{k=1}^{d} \frac{1}{d} \mathbb{E}_{p \sim \pi}\left[(p^{(k)})^2 \right] = \frac{1}{3}.$$

Therefore,

$$\begin{aligned}\mathbb{E}_{p\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_p^{\otimes n}}\left[\left\langle\hat{\theta},\sum_{i\in[n]}(Z_i-\mu_p)\right\rangle\right] &= 2\mathbb{E}_{p\sim\pi}\left[\left\langle\mathbb{E}_{S_n\sim\mathcal{D}_p^{\otimes n}}[\hat{\theta}],\mu_p\right\rangle\right]\\ &\geq \frac{1}{3}-2\varepsilon-3\delta,\end{aligned}$$

as was to be shown.

E.3. Proof of Theorem 4.2

Fix a learning algorithm $\mathcal{A}$, and let $\mathcal{D}$ be a distribution satisfies Lemma 7.3. Also, consider the structure introduced in the definition of CMI in Definition 3.2 and let $\tilde{\boldsymbol{Z}}=(Z_{j,i})_{j\in\{0,1\},i\in[n]}\sim\mathcal{D}^{\otimes(2\times n)}$. Let $\beta=1/12$ be a constant. Define the following set

$$\mathcal{I}=\left\{(i,j)\in[n]\times\{0,1\}\middle|\left\langle\hat{\theta}-\mu,Z_{j,i}-\mu\right\rangle\geq\beta/n \text{ and } \left\langle\hat{\theta}-\mu,Z_{1-j,i}-\mu\right\rangle<\beta/n\right\}.$$

Intuitively, $\mathcal{I}$ includes the subset of columns of supersample such that one of the samples has a *large* correlation to the output of the algorithm and the other one has *small* correlation to the output of the algorithm. Also, define the following event

$$\mathcal{G}=\left\{\forall i\in[n]:\left\langle\hat{\theta}-\mu,Z_{\bar{U}_i,i}-\mu\right\rangle<\beta/n\right\},$$

where $\bar{U}_i=1-U_i$. Intuitively, under the event $\mathcal{G}$ the correlation of the output and the *heldout samples* are insignificant. We can write

$$\begin{aligned}\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)&=\mathrm{H}(U|\tilde{\boldsymbol{Z}})-\mathrm{H}(U|\tilde{\boldsymbol{Z}},\hat{\theta})\\&=\mathrm{H}(U)-\mathrm{H}(U|\tilde{\boldsymbol{Z}},\hat{\theta})\\&=n-\mathrm{H}(U|\tilde{\boldsymbol{Z}},\hat{\theta}).\end{aligned}$$

where the last two steps follows from $U\perp\!\!\!\perp\tilde{\boldsymbol{Z}}$ and $\mathrm{H}(U)=n$. Notice that $\mathcal{I}$ is a $(\hat{\theta},\tilde{\boldsymbol{Z}})$-measurable random variable, thus, $\mathrm{H}(U|\tilde{\boldsymbol{Z}},\hat{\theta})=\mathrm{H}(U|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I})$. Define $\mathcal{I}^{(1)}$ as follows: $i\in\mathcal{I}^{(1)}$ iff $\exists j\in\{0,1\}$ such that $(i,j)\in\mathcal{I}$. Using this notation, we can write

$$\begin{aligned}\mathrm{H}(U|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I})&=\mathrm{H}(U_{\mathcal{I}^{(1)}},U_{(\mathcal{I}^{(1)})^c}|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I})\\&\leq\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I})+\mathrm{H}(U_{(\mathcal{I}^{(1)})^c}|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I}),\end{aligned}\tag{15}$$

where the last step follows from sub-additivity of Entropy. The second term in Equation (15) can be bounded by

$$\mathrm{H}(U_{(\mathcal{I}^{(1)})^c}|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I})\leq\mathbb{E}\left[(n-|\mathcal{I}|)\right].$$

Define the random variable $\hat{U}\in\{0,1\}^n$ as follows: for every $(i,j)\in\mathcal{I}$, let $\hat{U}_i=j$. For the remaining coordinates set $\hat{U}_i=0$. Notice that $\hat{U}$ is a $\mathcal{I}$ measurable random variable. Therefore, $\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I})=\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I},\hat{U})$. Then, we invoke Fano's inequality from Lemma B.1 to write

$$\begin{aligned}\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I},\hat{U})&\leq\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\hat{U})\\&\leq1+\mathrm{H}(U_{\mathcal{I}^c})\mathbb{P}\left(\{\exists(i,j)\in\mathcal{I}:U_i\neq j\}\right)\\&\leq1+n\mathbb{P}\left(\{\exists(i,j)\in\mathcal{I}:U_i\neq j\}\right).\end{aligned}$$

We claim that $\mathbb{P}\left(\{\exists(i,j)\in\mathcal{I}:U_i\neq j\}\right)\leq\mathbb{P}\left(\mathcal{G}^c\right)$. The proof is as follows: If there exists $(i,j)\in\mathcal{I}$ such that $U_i\neq j$, then, we have

$$\left\langle\hat{\theta}-\mu,Z_{\bar{U}_i,i}-\mu\right\rangle\geq\frac{\beta}{n},$$

by the definition of $\mathcal{I}$. Therefore, we conclude $\mathrm{H}(U_{\mathcal{I}^{(1)}}\big|\tilde{\boldsymbol{Z}},\hat{\theta},\mathcal{I})\leq1+n\mathbb{P}\left(\mathcal{G}^c\right)$. The conditional entropy can be upper-bounded by

$$\mathrm{H}(U\big|\tilde{\boldsymbol{Z}},\hat{\theta})\leq n-\mathbb{E}\left[|\mathcal{I}|\right]+1+n\mathbb{P}\left(\mathcal{G}^c\right).$$

By the definition of mutual information, we can lower bound $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ as follows

$$\begin{aligned}\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= n - \mathrm{H}(U|\tilde{\boldsymbol{Z}}, \hat{\theta}) \\ &= \mathbb{E}\left[|\mathcal{I}|\right] - 1 - n\mathbb{P}\left(\mathcal{G}^c\right).\end{aligned} \tag{16}$$

In the next step of the proof, we provide a lower bound on $\mathbb{E}\left[|\mathcal{I}|\right]$. Let us define a random variable that measures the *correlation* between the output and the i-th training samples:

$$c_i \triangleq \left\langle \hat{\theta} - \mu, Z_{U_i,i} - \mu \right\rangle.$$

Under the event $\mathcal{G}$, using Lemma B.4 we can lower bound $\mathbb{E}[|\mathcal{I}|]$ as follows

$$\begin{aligned}\mathbb{E}[|\mathcal{I}|] &\geq \mathbb{E}\left[|\mathcal{I}|\mathbb{1}\left[\mathcal{G}\right]\right] \\ &\geq \mathbb{E}\left[\left|\left\{i \in [n] : c_i \geq \frac{\beta}{n}\right\}\right|\mathbb{1}\left[\mathcal{G}\right]\right] \\ &\geq \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} c_i - \beta, 0\}\right)^2}{\sum_{i\in[n]} c_i^2}\mathbb{1}\left[\mathcal{G}\right]\right].\end{aligned} \tag{17}$$

Also, define the following event

$$\mathcal{E} \triangleq \mathcal{G} \cap \left\{\left\|\hat{\theta} - \mu\right\|^2 \leq \varepsilon\right\} \cap \left\{\sum_{i\in[n]} c_i^2 \leq 6\left\|\hat{\theta} - \mu\right\|^2\right\}.$$

Since $\mathcal{E} \subseteq \mathcal{G}$, we have

$$\begin{aligned}\mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} c_i - \beta, 0\}\right)^2}{\sum_{i\in[n]} c_i^2}\mathbb{1}\left[\mathcal{G}\right]\right] &\geq \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} c_i - \beta, 0\}\right)^2}{\sum_{i\in[n]} c_i^2}\mathbb{1}\left[\mathcal{E}\right]\right] \\ &\geq \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} c_i - \beta, 0\}\right)^2}{6\varepsilon}\mathbb{1}\left[\mathcal{E}\right]\right] \\ &= \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} c_i - \beta, 0\}\right)^2}{6\varepsilon}\right] - \mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} c_i - \beta, 0\}\right)^2}{6\varepsilon}\mathbb{1}\left[\mathcal{E}^c\right]\right],\end{aligned} \tag{18}$$

where the second step follows because under event $\mathcal{G}$, $\sum_{i\in[n]} c_i^2 \leq 6\left\|\hat{\theta} - \mu\right\|^2$ and $\left\|\hat{\theta} - \mu\right\|^2 \leq \varepsilon$. By convexity of $h_1(x) = x^2$ and $h_2(x) = \max\{x, 0\}$, we can use Jensen's inequality to obtain

$$\begin{aligned}\mathbb{E}\left[\frac{\left(\max\left\{\sum_{i\in[n]} c_i - \beta, 0\right\}\right)^2}{\varepsilon}\right] &\geq \frac{\left(\max\left\{\mathbb{E}\left[\sum_{i\in[n]} c_i\right] - \beta, 0\right\}\right)^2}{6\varepsilon} \\ &\geq \frac{\left(\frac{1}{3} - 2\varepsilon - 3\delta - \beta\right)^2}{6\varepsilon},\end{aligned}$$

where the last step follows from Lemma 7.3. Notice that by setting $\varepsilon < 1/24$ and $\delta < 1/48$, we have $\left(\frac{1}{3} - 2\varepsilon - 3\delta - \beta\right)^2 \geq 1/36$.

To upperbound the second term in Equation (18), first, notice that the following holds with probability one

$$\begin{aligned}\frac{\left(\max\{\sum_{i\in[n]} c_i - \beta, 0\}\right)^2}{3\varepsilon} &\leq \frac{\left(\sum_{i\in[n]} c_i\right)^2 + 2\beta^2}{6\varepsilon}\\ &\leq \frac{2\beta^2 + 16n^2}{6\varepsilon},\end{aligned}$$

where the last step follows from

$$\begin{aligned}\left\|\sum_{i\in[n]} c_i\right\| &= \left\|\left\langle \hat{\theta} - \mu, \sum_{i=1}^{n} (Z_{U_i,i} - \mu)\right\rangle\right\|\\ &\leq \left\|\hat{\theta} - \mu\right\| \left\|\sum_{i=1}^{n} (Z_{U_i,i} - \mu)\right\|\\ &\leq 4n.\end{aligned}$$

Then, in the next step we provide an upperbound on $\mathbb{P}(\mathcal{E}^c)$. Union bound implies that

$$\mathbb{P}\left(\mathcal{E}^c\right) \leq \mathbb{P}\left(\mathcal{G}^c\right) + \mathbb{P}\left(\left\|\hat{\theta} - \mu\right\|^2 > \varepsilon\right) + \mathbb{P}\left(\sum_{i\in[n]} c_i^2 > 6\left\|\hat{\theta} - \mu\right\|^2\right).$$

We want to set the parameters so that for a sufficiently small γ the following hold

$$\mathbb{P}\left(\mathcal{G}^c\right) \leq \gamma/n^2, \mathbb{P}\left(\left\|\hat{\theta} - \mu\right\|^2 > \varepsilon\right) \leq \gamma/n^2, \mathbb{P}\left(\sum_{i\in[n]} c_i^2 > 6\left\|\hat{\theta} - \mu\right\|^2\right) \leq \gamma/n^2. \tag{19}$$

Notice that

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}^c\right) &= \mathbb{P}\left(\max_{i\in[n]}\left\{\left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu\right\rangle\right\} \geq \beta/n\right)\\ &= \mathbb{E}\left[\mathbb{P}\left(\left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu\right\rangle \geq \beta/n \middle| U, \hat{\theta}\right)\right].\end{aligned}$$

By the construction of CMI in Definition 3.2, conditioned on U, $\hat{\theta}$, $Z_{\bar{U}_i,i}$ is i.i.d.from $\mathcal{D}$ for $i \in [n]$. Therefore, we can use Lemma B.8,

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}^c\right) &\leq n\mathbb{P}\left(\left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu\right\rangle \geq \beta/n\right)\\ &\leq n\exp\left(-\frac{d}{8n^2}\right),\end{aligned}$$

We can see setting $d \geq O(n^2\log(n^2))$, we have $\mathbb{P}\left(\mathcal{G}^c\right) \leq \gamma/n^2$ in Equation (19). Then, by the fact that $\mathcal{A}$ ε-learns $\mathcal{P}_{\text{scvx}}^{(d)}$ and Lemma 7.1 we have

$$\mathbb{P}\left(\left\|\hat{\theta} - \mu\right\|^2 > \varepsilon\right) \leq \delta = O\left(1/n^2\right).$$

Also, by Lemma B.7, given that $d > O\left(n^4\log(n)\right)$

$$\begin{aligned}\mathbb{P}\left(\sum_{i\in[n]} c_i^2 > 6\left\|\hat{\theta} - \mu\right\|\right) &= \mathbb{P}\left(\sum_{i\in[n]}\left\langle \hat{\theta} - \mu, Z_{U_i,i} - \mu\right\rangle^2 > 6\left\|\hat{\theta} - \mu\right\|\right)\\ &= \mathbb{E}\left[\mathbb{P}\left(\sum_{i\in[n]}\left\langle \hat{\theta} - \mu, Z_{U_i,i} - \mu\right\rangle^2 > 6\left\|\hat{\theta} - \mu\right\| \middle| U\right)\right]\\ &\leq O\left(1/n^2\right).\end{aligned}$$

In summary, we conclude that we can set the parameters such that in Equation (18)

$$\mathbb{E}\left[\frac{\left(\max\{\sum_{i\in[n]} c_i-\beta,0\}\right)^2}{3\varepsilon}\mathbb{1}\left[\mathcal{E}^c\right]\right]\leq\frac{2\beta^2+16n^2}{3\varepsilon}\mathbb{P}\left(\mathcal{G}^c\right)\leq O\left(\frac{1}{\varepsilon}\right).$$

Ergo, we have

$$\mathbb{E}\left[|\mathcal{I}|\mathbb{1}\left[\mathcal{G}\right]\right]\geq\Omega\left(\frac{1}{\varepsilon}\right).$$

E.4. Proof of Theorem 7.4

The algorithm is based on subsampling a subset of training samples to create a new dataset and feeding it into an empirical risk minimizer. Let $0<m\leq n$ be constants to be determined later. Let the training set $S_n=(Z_1,\dots,Z_n)$. The output of the algorithm $\hat{\theta}=\mathcal{A}_n(S)$ is

$$\hat{\theta}=\underset{\theta\in\Theta}{\arg\min}\Big\{\sum_{i\in[m]}f(\theta,Z_i)\Big\}.$$

Notice that $\hat{\theta}$ is unique since $f(\cdot,z)$ is a strongly convex function. By [SSSS09, Thm. 6], we have

$$\mathbb{E}[\mathrm{F}_{\mathcal{D}}(\hat{\theta})]-\min_{\theta\in\Theta}\mathrm{F}_{\mathcal{D}}(\hat{\theta})\leq\frac{4L^2}{\mu m}.$$

Since $\mathcal{A}$ is a function of the first m samples only, using the same argument as in the proof of Theorem 6.5, we can show that $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)\leq m$. Finally, setting $m=\frac{4L^2}{\mu\varepsilon}$ concludes the proof.

E.5. Corollaries of Proof of Theorem 4.2

Corollary E.1. *Let $\mathcal{P}^{(d)}_{scvx}$ be the problem instance described in Section 7.1.1. For every $\varepsilon<1/24$ and $\delta<1/48$ and for every ε-learner ($\mathcal{A}=\{\mathcal{A}_n\}_{n\in\mathbb{N}}$), with sample complexity $N(\cdot,\cdot)$ the following holds: for every $n\geq N(\varepsilon,\delta)$, $\delta<O(1/n^2)$, and $d\geq O(n^4\log(n))$, there exists a data distribution $\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})$ such that*

$$\mathbb{E}\left[\Big|\Big\{i\in[n]:\Big\langle\hat{\theta}-\mu,Z_i-\mu\Big\rangle\Big\}\geq\frac{1}{12n}\Big|\right]=\Omega\left(\frac{1}{\varepsilon}\right),$$

where $S_n=(Z_1,\dots,Z_n)\sim\mathcal{D}^{\otimes n}$, $\hat{\theta}=\mathcal{A}(S_n)$, and $\mu=\mathbb{E}_{Z\sim\mathcal{D}}\left[Z\right]$.

Corollary E.2. *Consider the structure introduced in the definition of CMI in Definition 3.2. Then, define the random variable*

$$\mathcal{I}=\Big\{(i,j)\in[n]\times\{0,1\}\Big|\,\langle\mathcal{A}_n(S_n)-\mu,Z_{j,i}-\mu\rangle\geq1/(12n)\text{ and }\langle\mathcal{A}_n(S_n)-\mu,Z_{1-j,i}-\mu\rangle<1/(12n)\Big\}.$$

Let $\mathcal{P}^{(d)}_{scvx}$ be the problem instance described in Section 7.1.1. For every $\varepsilon<1/24$ and $\delta<1/48$ and for every ε-learner ($\mathcal{A}=\{\mathcal{A}_n\}_{n\in\mathbb{N}}$), with sample complexity $N(\cdot,\cdot)$ the following holds: for every $n\geq N(\varepsilon,\delta)$, $\delta<O(1/n^2)$, and $d\geq O(n^4\log(n))$, there exists a data distribution $\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})$ such that

$$\mathbb{E}\left[|\mathcal{I}|\right]=\Omega\left(\frac{1}{\varepsilon}\right).$$

F. Proof of Memorization Results

F.1. Adversary Strategy

We describe the proposed strategy for the adversary in Algorithm 1 and Algorithm 2.

Algorithm 1 $\mathcal{Q}_{\text{cvx}}$: Attacker for Convex Losses

```
Inputs: $\hat{\theta} \in \Theta$, $Z \in \mathcal{Z}$, $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$.
$\mu = (\mu^{(1)}, \ldots, \mu^{(k)}) = \mathbb{E}_{Z \sim \mathcal{D}}[Z]$.
$A = \operatorname{diag}\left[\left\{\frac{144\varepsilon^2 - d(\mu^{(k)})^2}{1 - d(\mu^{(k)})^2}\right\}_{k=1}^{d}\right]$.
$\beta = \varepsilon$.
if $\left\langle \hat{\theta}, A\,(Z - \mu) \right\rangle \geq \beta/n$ then
    $\hat{b} = 1$
else
    $\hat{b} = 0$
Output $\hat{b}$
```

Algorithm 2 $\mathcal{Q}_{\text{scvx}}$: Adversary for Strongly Convex Losses

```
Inputs: $\hat{\theta} \in \Theta$, $Z \in \mathcal{Z}$, $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$.
$\mu = \mathbb{E}_{Z \sim \mathcal{D}}[Z]$
$\beta = 1/12$.
if $\left\langle \hat{\theta} - \mu, Z - \mu \right\rangle \geq \frac{\beta}{4n}$ then
    $\hat{b} = 1$
else
    $\hat{b} = 0$
Output $\hat{b}$
```

Algorithm 3 $\mathsf{FP}_{\text{scvx}}$: Fingerprint detector for Strongly Convex Losses

```
Inputs: $\hat{\theta} \in \Theta$, $(Z_0, \ldots, Z_n) \in \mathcal{Z}^{n+1}$, $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$.
$\mu = \mathbb{E}_{Z \sim \mathcal{D}}[Z]$
$\beta = 1/12$.
$\mathcal{B}_{\text{FP}} = \varnothing$
for $i \in \{0, \ldots, n\}$ do:
    if $\left\langle \hat{\theta} - \mu, Z_i - \mu \right\rangle \geq \frac{\beta}{n}$ then
        $\mathcal{B}_{\text{FP}} = \mathcal{B}_{\text{FP}} \cup \{i\}$
Output $\mathcal{B}_{\text{FP}}$
```

Algorithm 4 $\mathsf{CR}_{\text{scvx}}$: Correlation-Reduction for Strongly Convex Losses

```
Inputs: $\hat{\theta} \in \Theta$, $(Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, $Z_0 \sim \mathcal{D}$
$\tilde{\mu} = Z_0$
$\beta = 1/12$.
$\mathcal{B}_{\text{corr-red}} = \varnothing$
$w = \hat{\theta}$
for $i \in [n]$ do:
    if $\left\langle \hat{\theta}, Z_i - \tilde{\mu} \right\rangle \geq \frac{\beta}{2n}$ then
        $\mathcal{B}_{\text{corr-red}} = \mathcal{B}_{\text{corr-red}} \cup \{i\}$
        if $\left|\mathcal{B}_{\text{corr-red}}\right| = \frac{2}{\varepsilon} \log\left(\frac{1}{\delta}\right)$ then
            Sample $\mathcal{R} \subseteq [n]$ a uniform random subset of size $\frac{2}{\varepsilon} \log(\frac{1}{\delta})$ from $[n]$
            $w = \mu_{\text{emp}}(\mathcal{R})$
            Break
Output $w, \mathcal{B}_{\text{corr-red}}$
```

F.2. Proof of Theorem 4.5

Let $b = (b_1, \dots, b_n)$ denote the outcome of fair coin at each round of the game described in Definition 4.3. Then, let $\hat{b}_i = \mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{b_i,i}, \mathcal{D}\right)$ for each round $i \in [n]$ and let us denote the output of the adversary as $(\hat{b}_1, \dots, \hat{b}_n) \in \{0,1\}^n$.

F.2.1. Soundness Analysis

Define the following event

$$\mathcal{G} = \left\{\forall i \in [n] : \left\langle \hat{\theta}, A(Z_{0,i} - \mu)\right\rangle < \beta/n\right\}.$$

Notice that

$$\begin{aligned}
&\mathbb{P}\left(\exists i \in [n] \colon \mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1\right)\\
&= \mathbb{P}\left(\left\{\, \exists i \in [n] \colon \mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\right\} \,\cap\, \mathcal{G}\right) + \mathbb{P}\left(\left\{\, \exists i \in [n] \colon \mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\right\} \,\cap\, \mathcal{G}^c\right)\\
&\leq \mathbb{P}\left(\left\{\, \exists i \in [n] \colon \mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\right\} \,\cap\, \mathcal{G}\right) + \mathbb{P}\left(\mathcal{G}^c\right).
\end{aligned}$$

We claim that $\mathbb{P}\left(\left\{\, \exists i \in [n] : \mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\right\} \,\cap\, \mathcal{G}\right) = 0$. It follows from the following observation: $\exists i \in [n] :$ $\mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1$ can happen if and only if there exists $i \in [n]$ such that $\left\langle \hat{\theta}, A\left(Z_{0,i} - \mu\right)\right\rangle \geq \beta/n$. However, the intersection of this event with $\mathcal{G}$ is empty by the definition of $\mathcal{G}$. Therefore, we can write

$$\mathbb{P}\left(\exists i \in [n] \colon \mathcal{Q}_{\text{cvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1\right) \leq \mathbb{P}\left(\mathcal{G}^c\right).$$

To upperbound $\mathbb{P}\left(\mathcal{G}^c\right)$, notice

$$\mathbb{P}\left(\mathcal{G}^c\right) = \mathbb{P}\left(\exists i \in [n] : \left\langle \hat{\theta}, A(Z_{0,i} - \mu)\right\rangle \geq \beta/n\right).$$

By the fact that $\hat{\theta} \perp\!\!\!\perp Z_{0,i}$ for every $i \in [n]$, we can use Lemma B.8 and the fact that $\left\|\hat{\theta}\right\| \leq 1$, to write

$$\mathbb{P}\left(\exists i \in [n] : \left\langle A\hat{\theta}, Z_{0,i} - \mu\right\rangle \geq \beta/n\right) \leq n \exp\left(-\frac{d}{2n^2\varepsilon^2}\right) \leq \xi,$$

given $d \geq \Omega(n^2 \log(n/\xi))$. Notice that by assumption $\varepsilon < 1$. This concludes the soundness analysis.

F.2.2. Recall Analysis

The construction of hard problem instance is given in Section 6.1.1. Let $\mathcal{A}$ be an arbitrary ε-learner and let $\mathcal{D}$ be a distribution satisfies Lemma 6.3. Let us define

$$\mathcal{I} = \{i \in [n] : \left\langle \hat{\theta}, A(Z_{1,i} - \mu)\right\rangle \geq \beta/n\}$$

where β is given in Algorithm 1. As seen in Equation (17) in the proof of Theorem 4.1, we have

$$\mathbb{E}\left[|\mathcal{I}|\right] = \Omega(1/\varepsilon^2).$$

An important observation is that $\sum_{i=1}^{n} \mathbb{1}\left[\mathcal{Q}\left(\hat{\theta}, Z_{1,i}, \mathcal{D}\right)\right] = |\mathcal{I}|$. Then, notice the total number of samples is $\tilde{O}(1/\varepsilon^2)$. Therefore, using reverse Markov's inequality in Lemma B.9, we obtain

$$\mathbb{P}(|\mathcal{I}| \geq \Omega(1/\varepsilon^2)) \geq p_0,$$

where p_0 is a constant independent of n and ε.

F.3. Proof of Theorem 4.6

Let $b = (b_1, \dots, b_n)$ denote the outcome of fair coin at each round of the game described in Definition 4.3. Then, let $\hat{b}_i = \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{b_i,i}, \mathcal{D}\right)$ for each round $i \in [n]$ and let us denote the output of the adversary as $(\hat{b}_1, \dots, \hat{b}_n) \in \{0,1\}^n$.

F.3.1. Soundness Analysis

Define the following event

$$\mathcal{G} = \left\{\forall i \in [n] : \left\langle \hat{\theta} - \mu, Z_{0,i} - \mu \right\rangle < \beta/n \right\}.$$

Notice that

$$\begin{aligned}
&\mathbb{P}\left(\exists i \in [n] \colon \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1\right) \\
&= \mathbb{P}\left(\{\, \exists i \in [n] \colon \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\} \cap \mathcal{G}\right) + \mathbb{P}\left(\{\, \exists i \in [n] \colon \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\} \cap \mathcal{G}^c\right) \\
&\le \mathbb{P}\left(\{\, \exists i \in [n] \colon \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\} \cap \mathcal{G}\right) + \mathbb{P}\left(\mathcal{G}^c\right).
\end{aligned}$$

We claim that $\mathbb{P}\left(\{\, \exists i \in [n] : \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1 \,\} \cap \mathcal{G}\right) = 0$. It follows from the following observation: $\mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1$ can happen if and only if $\left\langle \hat{\theta} - \mu, Z_{0,i} - \mu \right\rangle \ge \beta/n$. However, the intersection of this event with $\mathcal{G}$ is empty by the definition of $\mathcal{G}$. Therefore, we can write

$$\mathbb{P}\left(\exists i \in [n] \colon \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1\right) \le \mathbb{P}\left(\mathcal{G}^c\right).$$

Since $Z_{0,i} \perp\!\!\!\perp \hat{\theta}$, we can use Lemma B.8 to write

$$\mathbb{P}\left(\mathcal{G}^c\right) \le n \exp\left(-\frac{d}{4n^2}\right).$$

By setting $d \ge \Omega(n^2 \log(n/\xi))$, we obtain that

$$\mathbb{P}\left(\exists i \in [n] \colon \mathcal{Q}_{\text{scvx}}\left(\hat{\theta}, Z_{0,i}, \mathcal{D}\right) = 1\right) \le \xi.$$

This concludes the soundness analysis.

F.3.2. Recall Analysis

The construction of the hard problem instance is given in Section 7.1.1. Let $\mathcal{A}$ be an arbitrary ε-learner and let $\mathcal{D}$ be a distribution that satisfies Lemma 7.3. Consider the algorithms given in Algorithms 2 to 4 and using them define the following random variables:

$$\begin{aligned}
(Z_0, Z_1, \ldots, Z_n) &\sim \mathcal{D}^{\otimes(n+1)}, \\
\hat{\theta} &= \mathcal{A}_n(Z_1, \ldots, Z_n), \\
\mathcal{B}_{\text{adversary}} &= \{i \in [n] : \hat{b}_i = 1\}, \\
w, \mathcal{B}_{\text{corr-red}} &= \mathsf{CR}_{\text{scvx}}\left(\hat{\theta}, (Z_1, \ldots, Z_n), Z_0\right), \\
\mathcal{B}_{\text{FP}} &= \mathsf{FP}_{\text{scvx}}\left(w, (Z_0, \ldots, Z_n), \mathcal{D}\right).
\end{aligned}$$

In particular, Z_0 is a sample drawn from $\mathcal{D}$ which is independent of the training set, i.e., $(Z_1, \ldots, Z_n)$.

Recall that our goal is to show that $\mathbb{P}\left(\left|\mathcal{B}_{\text{adversary}}\right| = \Omega(1/\varepsilon)\right)$ is greater than a universal constant. Our approach is as follows: In the first step, we show that, with a high probability, $\mathcal{B}_{\text{corr-red}} \subseteq \mathcal{B}_{\text{adversary}}$. Then, in the second step, we will show that with a high probability $|\mathcal{B}_{\text{FP}}| \le |\mathcal{B}_{\text{corr-red}}| + 1$. In the third step, we will show that $\mathbb{E}\left[\left|\mathcal{B}_{\text{FP}}\right|\right] = \Omega(1/\varepsilon)$ which gives us $\mathbb{E}\left[\left|\mathcal{B}_{\text{corr-red}}\right|\right] = \Omega(1/\varepsilon)$. Finally, by the fact that $\mathcal{B}_{\text{corr-red}} = O\left(1/\varepsilon\right)$ with probability one and reverse Markov's inequality, we obtain that $\mathbb{P}(\left|\mathcal{B}_{\text{corr-red}}\right| \ge \Omega(1/\varepsilon))$ is greater than a universal constant. Combining this result with Step 1, concludes the proof.

Step 1: with a high probability, $\mathcal{B}_{\text{corr-red}} \subseteq \mathcal{B}_{\text{adversary}}$. Simple calculations show that

$$\left\langle \hat{\theta}, Z_i - \tilde{\mu} \right\rangle = \left\langle \hat{\theta} - \mu, Z_i - \mu \right\rangle + \left\langle \hat{\theta}, \mu - \tilde{\mu} \right\rangle + \left\langle \mu, Z_i - \mu \right\rangle.$$

Then, we can write

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}_{\text{corr-red}} \not\subseteq \mathcal{B}_{\text{adversary}}\right) &\leq \mathbb{P}\left(\exists i \in [n] : \left\langle \hat{\theta}, Z_i - \tilde{\mu} \right\rangle \geq \frac{\beta}{2n} \wedge \left\langle \hat{\theta} - \mu, Z_i - \mu \right\rangle < \frac{\beta}{4n}\right) \\
&\leq \mathbb{P}\left(\exists i \in [n] : \left\langle \hat{\theta}, \mu - \tilde{\mu} \right\rangle + \left\langle \mu, Z_i - \mu \right\rangle \geq \frac{\beta}{4n}\right) \\
&\leq \mathbb{P}\left(\left\langle \hat{\theta}, \mu - \tilde{\mu} \right\rangle \geq \frac{\beta}{4n}\right) + \mathbb{P}\left(\exists i \in [n] : \left\langle \mu, Z_i - \mu \right\rangle \geq \frac{\beta}{4n}\right) \\
&\leq \exp\left(-\frac{d \cdot \beta^2}{32 n^2 \left\|\mu\right\|^2}\right) + n \exp\left(-\frac{d\beta^2}{32 n^2 \left\|\mu\right\|^2}\right) \\
&\leq (n+1) \exp\left(-\frac{d \cdot \beta^2}{32 n^2}\right),
\end{aligned}$$

where the second step follows from union bound and the third step follows from Lemma B.8. This shows that setting $d = \Omega\left(n^2 \log(n^2)\right)$, we obtain

$$\mathbb{P}\left(\exists i \in [n] : \left\langle \hat{\theta}, Z_i - \tilde{\mu} \right\rangle \geq \frac{\beta}{2n} \wedge \left\langle \hat{\theta} - \mu, Z_i - \mu \right\rangle < \frac{\beta}{4n}\right) \leq O\left(\frac{1}{n}\right).$$

This is equivalent to

$$\mathbb{P}\left(\mathcal{B}_{\text{corr-red}} \subseteq \mathcal{B}_{\text{adversary}}\right) \geq 1 - O\left(\frac{1}{n}\right).$$

Step 2: with a high probability, $|\mathcal{B}_{\mathbf{FP}}| \leq |\mathcal{B}_{\mathbf{corr\text{-}red}}| + 1$. Notice that $\left|\mathcal{B}_{\text{FP}}\right| = \left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| + \left|\mathcal{B}_{\text{FP}} \cap \{0\}\right|$. We can write

$$\begin{aligned}
\mathbb{P}\left(\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \left|\mathcal{B}_{\text{corr-red}}\right|\right) &= \mathbb{P}\left(\left\{\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \left|\mathcal{B}_{\text{corr-red}}\right|\right\} \wedge \{w = \hat{\theta}\}\right) \\
&\quad + \mathbb{P}\left(\left\{\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \left|\mathcal{B}_{\text{corr-red}}\right|\right\} \wedge \{w = \theta_0\}\right).
\end{aligned} \tag{20}$$

For the first term in Equation (20),

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \left|\mathcal{B}_{\text{corr-red}}\right|\right\} \wedge \{w = \hat{\theta}\}\right) \\
&\leq \mathbb{P}\left(\left\{\exists i \in \{1, \ldots, n\} : i \in \mathcal{B}_{\text{FP}} \wedge i \notin \mathcal{B}_{\text{corr-red}}\right\} \wedge \{w = \hat{\theta}\}\right) \\
&= \mathbb{P}\left(\left\{\exists i \in \{1, \ldots, n\} : \left\langle \hat{\theta} - \mu, Z_i - \mu \right\rangle \geq \frac{\beta}{n} \wedge \left\langle \hat{\theta}, Z_i - \tilde{\mu} \right\rangle < \frac{\beta}{2n}\right\} \wedge \{w = \hat{\theta}\}\right) \\
&\leq \mathbb{P}\left(\left\{\exists i \in \{1, \ldots, n\} : \left\langle \hat{\theta} - \mu, Z_i - \mu \right\rangle \geq \frac{\beta}{n} \wedge \left\langle \hat{\theta}, Z_i - \tilde{\mu} \right\rangle < \frac{\beta}{2n}\right\}\right) \\
&\leq O\left(\frac{1}{n}\right),
\end{aligned}$$

where the last step follows from Lemma B.8. Then, for the second term in Equation (20),

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \left|\mathcal{B}_{\text{corr-red}}\right|\right\} \wedge \{w = \mu_{\text{emp}}(\mathcal{R})\}\right) \\
&= \mathbb{P}\left(\left\{\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \frac{2}{\varepsilon} \log(1/\delta)\right\} \wedge \{w = \mu_{\text{emp}}(\mathcal{R})\}\right),
\end{aligned}$$

where the last line follows because under the event $w = \mu_{\text{emp}}(\mathcal{R})$, $\left|\mathcal{B}_{\text{corr-red}}\right| = \frac{2}{\varepsilon} \log(1/\delta)$. Notice that $|\mathcal{R}| = \frac{2}{\varepsilon} \log(1/\delta)$ and $\mathcal{R}$ is independent of every other random variable. Therefore, the event $\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \frac{2}{\varepsilon} \log(1/\delta)$ is a subset of the event that there exists $i \notin \mathcal{R}$ such that $\langle w - \mu, Z_i - \mu \rangle > \frac{\beta}{n}$. However, notice that $w \perp\!\!\!\perp Z_i$ by the description of Algorithm 4. Therefore, we can write

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left|\mathcal{B}_{\text{FP}} \cap \{1, \ldots, n\}\right| > \frac{2}{\varepsilon} \log(1/\delta)\right\} \wedge \{w = \mu_{\text{emp}}(\mathcal{R})\}\right) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\exists i \notin \mathcal{R} : \langle \mu_{\text{emp}}(\mathcal{R}) - \mu, Z_i - \mu \rangle \geq \frac{\beta}{n} \middle| \mathcal{R}\right)\right].
\end{aligned}$$

By an application of Lemma B.8, we have

$$\mathbb{P}\left(\exists i \notin \mathcal{R} : \langle \mu_{\text{emp}}(\mathcal{R}) - \mu, Z_i - \mu \rangle \geq \frac{\beta}{n} \middle| \mathcal{R}\right) \leq n \cdot \exp\left(-\frac{d\beta^2}{32n^2}\right).$$

It can be seen by setting $d = \Omega(n^2 \log(n^2)$, we obtain that this probability is at most $O(1/n)$. Therefore, combining these two upperbounds with Equation (20) shows that with probability at least $1 - O(1/n)$, we have

$$\begin{aligned}\left|\mathcal{B}_{\text{FP}}\right| &= \left|\mathcal{B}_{\text{FP}} \cap \{1, \dots, n\}\right| + \left|\mathcal{B}_{\text{FP}} \cap \{0\}\right| \\ &\leq \left|\mathcal{B}_{\text{corr-red}}\right| + 1,\end{aligned}$$

as was to be shown.

Step 3: $\mathbb{E}\left[\mathcal{B}_{\textbf{FP}}\right] = \Omega(1/\varepsilon)$ We claim that w (output of Algorithm 4) satisfies the definition of ε-learner. The reason is as follows: w can be either $\hat{\theta} = \mathcal{A}_n(S_n)$ and $\mu_{\text{emp}}(\mathcal{R})$. Notice that $\mathcal{A}_n$ is an ε-learner by assumption. Consider the case that $w = \mu_{\text{emp}}(\mathcal{R})$. Then, consider

$$\begin{aligned}\mathbb{P}\left(\left\|\mu_{\text{emp}}(\mathcal{R}) - \mu\right\|^2 > \varepsilon\right) &= \mathbb{E}\left[\mathbb{P}\left(\left\|\mu_{\text{emp}}(\mathcal{R}) - \mu\right\|^2 > \varepsilon \middle| \mathcal{R}\right)\right] \\ &\leq \delta,\end{aligned}$$

where the last step follows from Lemma B.5. Therefore, by a union bound we see that the output of Algorithm 4 has an excess error of ε, with probability at least $1 - 2\delta$ with the sample complexity of $N(\varepsilon, \delta)$ where N is the sample complexity of $\mathcal{A}$. In Corollary E.1 we showed that for every ε-learner, we have

$$\mathbb{E}\left[\left|\mathcal{B}_{\text{FP}}\right|\right] = \Omega\left(\frac{1}{\varepsilon}\right).$$

Step 4: Conclusion. First, we provide a lowerbound on the $\mathbb{E}[|\mathcal{B}_{\text{corr-red}}|]$ as follows

$$\begin{aligned}\mathbb{E}\left[\left|\mathcal{B}_{\text{corr-red}}\right|\right] &= \mathbb{E}\left[\left|\mathcal{B}_{\text{corr-red}}\right| \cdot \mathbb{1}\left[\left|\mathcal{B}_{\text{corr-red}}\right| + 1 \geq \left|\mathcal{B}_{\text{FP}}\right|\right]\right] + \mathbb{E}\left[\left|\mathcal{B}_{\text{corr-red}}\right| \cdot \mathbb{1}\left[\left|\mathcal{B}_{\text{corr-red}}\right| + 1 < \left|\mathcal{B}_{\text{FP}}\right|\right]\right] \\ &\geq \mathbb{E}\left[\left|\mathcal{B}_{\text{FP}}\right| \cdot \mathbb{1}\left[\left|\mathcal{B}_{\text{corr-red}}\right| + 1 \geq \left|\mathcal{B}_{\text{FP}}\right|\right]\right] - 1 + \mathbb{E}\left[\left|\mathcal{B}_{\text{corr-red}}\right| \cdot \mathbb{1}\left[\left|\mathcal{B}_{\text{corr-red}}\right| + 1 < \left|\mathcal{B}_{\text{FP}}\right|\right]\right] \\ &\geq \mathbb{E}\left[\left|\mathcal{B}_{\text{FP}}\right|\right] - 1 + \mathbb{E}\left[\left(\left|\mathcal{B}_{\text{corr-red}}\right| - \left|\mathcal{B}_{\text{FL}}\right|\right) \cdot \mathbb{1}\left[\left|\mathcal{B}_{\text{corr-red}}\right| + 1 < \left|\mathcal{B}_{\text{FP}}\right|\right]\right] \\ &\geq \mathbb{E}\left[\left|\mathcal{B}_{\text{FP}}\right|\right] - 1 - n\mathbb{P}\left(\left|\mathcal{B}_{\text{corr-red}}\right| + 1 < \left|\mathcal{B}_{\text{FP}}\right|\right) \\ &\geq \Omega\left(\frac{1}{\varepsilon}\right),\end{aligned}$$

where the last line follows from Step 2 and Step 3. Since with probability one $\left|\mathcal{B}_{\text{corr-red}}\right| \leq \frac{2}{\varepsilon}\log\left(\frac{1}{\delta}\right)$, reverse Markov's inequality from Lemma B.9 gives

$$\mathbb{P}\left(\left|\mathcal{B}_{\text{corr-red}}\right| = \Omega\left(\frac{1}{\varepsilon}\right)\right) \geq p_0,$$

where p_0 is a universal constant. Also, we showed in Step 1 that

$$\mathbb{P}\left(\left|\mathcal{B}_{\text{adversary}}\right| \geq \left|\mathcal{B}_{\text{corr-red}}\right|\right) \geq 1 - O(\frac{1}{n}).$$

Combining these two facts using union bound concludes the proof.

G. Proofs of Lowerbound for Individual-Sample CMI

In this part, we show that our proof techniques for Theorem 4.1 and Theorem 4.2 easily extend to ISCMI. First, we begin with the strong convex case. Let $\beta = 1/12$ and define

$$\mathcal{I} = \left\{(i,j) \in [n] \times \{0,1\} \middle| \left\langle \hat{\theta} - \mu, Z_{j,i} - \mu \right\rangle \geq \beta/n \text{ and } \left\langle \hat{\theta} - \mu, Z_{1-j,i} - \mu \right\rangle < \beta/n \right\}.$$

Also, define $\mathcal{I}^{(1)}$ as follows: $i \in \mathcal{I}^{(1)}$ iff $\exists j \in \{0,1\}$ such that $(i,j) \in \mathcal{I}$. We also introduce the following events

$$\mathcal{G}_i = \left\{ \left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu \right\rangle < \beta/n \right\}, \quad \mathcal{M}_i = \left\{ i \in \mathcal{I}^{(1)} \right\},$$

where $\bar{U}_i = 1 - U_i$. Then, we can write

$$\sum_{i=1}^{n} I(\hat{\theta}, U_i \big| Z_{0,i}, Z_{1,i}) = n - \sum_{i=1}^{n} \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}).$$

In the next step, for every $i \in [n]$, we provide an upperbound on $\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i})$. First, notice that $\mathbb{1}\left[\mathcal{M}_i\right]$ is a $\left(\hat{\theta}, Z_{0,i}, Z_{1,i}\right)$-measurable random variable. Therefore,

$$\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}) = \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right]). \tag{21}$$

Using the monotonicity and chain rule of entropy, we can write

$$\begin{aligned}
\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right]) &\leq \mathrm{H}(U_i, \mathbb{1}\left[\mathcal{G}_i\right] \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right]) \\
&= \mathrm{H}(\mathbb{1}\left[\mathcal{G}_i\right] \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right]) + \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right], \mathbb{1}\left[\mathcal{G}_i\right]) \\
&\leq \mathrm{H}(\mathbb{1}\left[\mathcal{G}_i\right]) + \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right], \mathbb{1}\left[\mathcal{G}_i\right]) \\
&= \mathrm{H}_b\left(\mathbb{P}\left(\mathcal{G}_i^c\right)\right) + \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right], \mathbb{1}\left[\mathcal{G}_i\right]),
\end{aligned}$$

where the third step follows because conditioning does not increase entropy and the last step follows because $\mathbb{1}\left[\mathcal{G}_i\right]$ is a binary random variable. Then, we can write

$$\begin{aligned}
\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{M}_i\right], \mathbb{1}\left[\mathcal{G}_i\right]) &= \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{G}_i\right], \mathbb{1}\left[\mathcal{M}_i\right] = 0)\mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 0\right) \\
&\quad + \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{G}_i\right] = 1, \mathbb{1}\left[\mathcal{M}_i\right] = 1)\mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 1 \wedge \mathbb{1}\left[\mathcal{G}_i\right] = 1\right) \\
&\quad + \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{G}_i\right] = 0, \mathbb{1}\left[\mathcal{M}_i\right] = 1)\mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 1 \wedge \mathbb{1}\left[\mathcal{G}_i\right] = 0\right).
\end{aligned}$$

We use the following estimates for each term. Since U_i is a bianry random variable, we have

$$\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{G}_i\right], \mathbb{1}\left[\mathcal{M}_i\right] = 0)\mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 0\right) \leq \mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 0\right).$$

Then, conditioned on $\mathbb{1}\left[\mathcal{G}_i\right] = \mathbb{1}\left[\mathcal{M}_i\right] = 1$, U_i is given by j where $(i,j) \in \mathcal{I}$ since

$$\begin{aligned}
\{(i,j) \in \mathcal{I}\} \cap \mathcal{G}_i &\Rightarrow \left\{ \left\langle \hat{\theta} - \mu, Z_{j,i} - \mu \right\rangle \geq \beta/n \text{ and } \left\langle \hat{\theta} - \mu, Z_{1-j,i} - \mu \right\rangle < \beta/n \right\} \cap \left\{ \left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu \right\rangle < \beta/n \right\} \\
&\Rightarrow \{j = U_i\}.
\end{aligned}$$

Therefore,

$$\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{G}_i\right] = 1, \mathbb{1}\left[\mathcal{M}_i\right] = 0) = 0.$$

For the last term, since U_i is a binary random variable, we can write

$$\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}, \mathbb{1}\left[\mathcal{G}_i\right] = 0, \mathbb{1}\left[\mathcal{M}_i\right] = 0)\mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 1 \wedge \mathbb{1}\left[\mathcal{G}_i\right] = 0\right) \leq \mathbb{P}\left(\mathbb{1}\left[\mathcal{G}_i\right] = 0\right).$$

In summary, we showed that

$$\mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}) \leq \mathbb{P}\left(\mathbb{1}\left[\mathcal{G}_i\right] = 0\right) + \mathrm{H}_b\left(\mathbb{P}\left(\mathcal{G}_i^c\right)\right) + \mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 0\right).$$

Using it, we can upperbound the sum of the conditional entropy as

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}) &\leq \sum_{i=1}^{n} \mathrm{H}_b\left(\mathbb{P}\left(\mathcal{G}_i^c\right)\right) + \mathbb{P}\left(\mathcal{G}_i^c\right) + \mathbb{P}\left(\mathbb{1}\left[\mathcal{M}_i\right] = 0\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}\left[\mathbb{1}\left[\mathcal{M}_i\right] = 0\right]\right] + \mathbb{P}\left(\mathcal{G}_i^c\right) + \mathrm{H}_b\left(\mathbb{P}\left(\mathcal{G}_i^c\right)\right) \\
&= \mathbb{E}\left[(n - \left|\mathcal{I}\right|)\right] + \sum_{i=1}^{n} \mathbb{P}\left(\mathcal{G}_i^c\right) + \sum_{i=1}^{n} \mathrm{H}_b\left(\mathbb{P}\left(\mathcal{G}_i^c\right)\right).
\end{aligned}$$

Next, we provide an estimate for $\mathbb{P}\left(\mathcal{G}_i^c\right)$

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}_i^c\right) &= \mathbb{P}\left(\left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu \right\rangle \geq \beta/n\right)\\ &= \mathbb{E}\left[\mathbb{P}\left(\left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu \right\rangle \geq \beta/n \middle| \hat{\theta}, U_i\right)\right].\end{aligned}$$

Since conditioned on U_i and $\hat{\theta}$, $Z_{\bar{U}_i,i} \sim \mathcal{D}$ and $\mathcal{D}$ is a product measure, using Lemma B.8, we have

$$\mathbb{P}\left(\left\langle \hat{\theta} - \mu, Z_{\bar{U}_i,i} - \mu \right\rangle \geq \beta/n\right) \leq O\left(\frac{1}{n^2}\right).$$

Also, by the well-known inequality, $\mathrm{H}_b(x) \leq -x\log(x) + x$ for $x \in [0,1]$, we have

$$\mathrm{H}_b\left(\mathbb{P}\left(\mathcal{G}_i^c\right)\right) \leq O\left(\frac{\log(n)}{n^2}\right).$$

Therefore,

$$\sum_{i=1}^{n} \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i}) \leq n - \mathbb{E}[|\mathcal{I}|] + O\left(\frac{\log(n)}{n}\right).$$

Plugging this upperbound into Equation (21),

$$\begin{aligned}\sum_{i=1}^{n} I(\hat{\theta}, U_i \big| Z_{0,i}, Z_{1,i}) &= n - \sum_{i=1}^{n} \mathrm{H}(U_i \big| \hat{\theta}, Z_{0,i}, Z_{1,i})\\ &\geq \mathbb{E}[|\mathcal{I}|] - O\left(\frac{\log(n)}{n}\right).\end{aligned}$$

Finally, we use Corollary E.2, to conclude that

$$\begin{aligned}\sum_{i=1}^{n} I(\hat{\theta}, U_i \big| Z_{0,i}, Z_{1,i}) &\geq \mathbb{E}[|\mathcal{I}|] - O\left(\frac{\log(n)}{n}\right)\\ &\geq \Omega\left(\frac{1}{\varepsilon}\right) - O\left(\frac{\log(n)}{n}\right)\\ &\geq \Omega\left(\frac{1}{\varepsilon}\right),\end{aligned}$$

where the last step follows since the minimum number of samples to ε-learn $\mathcal{P}_{\text{scvx}}^{(d)}$ is $n \geq \Omega(1/\varepsilon)$.

The proof of CLB subclass of SCOs is the same: using the same techniques we can lowerbound ISCMI by $\mathbb{E}[|\mathcal{I}|]$ and then by Corollary C.2 the result follows.