Proceedings of Machine Learning Research vol 201:1–44, 2023 34th International Conference on Algorithmic Learning Theory

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

Mahdi Haghifam[*] MAHDI.HAGHIFAM@MAIL.UTORONTO.CA
University of Toronto,Vector Institute

Borja Rodríguez-Gálvez[*] BORJARG@KTH.SE
KTH Royal Institute of Technology

Ragnar Thobaben RAGNART@KTH.SE
KTH Royal Institute of Technology

Mikael Skoglund SKOGLUND@KTH.SE
KTH Royal Institute of Technology

Daniel M. Roy DANIEL.ROY@UTORONTO.CA
University of Toronto,Vector Institute

Gintare Karolina Dziugaite GKDZ@GOOGLE.COM
Google Research, Mila, McGill

Editors: Shipra Agrawal and Francesco Orabona

Abstract

To date, no "information-theoretic" frameworks for reasoning about generalization error have been shown to establish minimax rates for gradient descent in the setting of stochastic convex optimization. In this work, we consider the prospect of establishing such rates via several existing information-theoretic frameworks: input-output mutual information bounds, conditional mutual information bounds and variants, PAC-Bayes bounds, and recent conditional variants thereof. We prove that none of these bounds are able to establish minimax rates. We then consider a common tactic employed in studying gradient methods, whereby the final iterate is corrupted by Gaussian noise, producing a noisy "surrogate" algorithm. We prove that minimax rates cannot be established via the analysis of such surrogates. Our results suggest that new ideas are required to analyze gradient descent using information-theoretic techniques.

1. Introduction

In this work, we uncover limitations of information-theoretic techniques towards analyzing stochastic gradient descent. To do so, we extend existing information-theoretic frameworks for reasoning about generalization to the setting of stochastic convex optimization (SCO) [**shalev2009stochastic**]. Despite the resulting bounds being provably tight, we develop an SCO problem in which the mutual information terms underlying these bounds are too large to demonstrate that subgradient methods [**cauchy1847methode**; **robbins1951stochastic**; **bubeck2015convex**] obtain minimax rates. We also consider the introduction of isotropic Gaussian noise to the final iterate and demonstrate a fundamental tradeoff between optimization error and expected generalization error that never yields minimax rates. Our results also cast doubt on the effectiveness of using isotropic Gaussian noise to study subgradient methods in other settings, such as deep learning.

[*] Mahdi Haghifam and Borja Rodríguez-Gálvez are equal-contribution authors.

Information-theoretic bounds are, by their nature, distribution- and algorithm-dependent. These bounds have shown some promises: for instance, these key properties enable information-theoretic frameworks to achieve numerically non-vacuous generalization guarantees for stochastic gradient Langevin dynamics (SGLD) with modern deep-learning datasets and architectures [**negrea2019information**; **haghifam2020sharpened**; **li2019generalization**; **pmlr-v162-banerjee22a**]. Therefore, it is natural to wonder whether the underlying quantity—mutual information—offers a potentially unifying tool to reason about generalization.

Information-theoretic techniques have long been used to classify the hardness of learning problems in terms of lower bounds on minimax risk. The development of information-theoretic techniques to upper bound minimax risk is a more recent approach. A stream of work has produced a variety of bounds on the generalization error of learning algorithms in terms of the (conditional) mutual information between the training data and certain statistics of the learned predictor. In the case of binary classification, the suprema of certain such bounds match (known) minimax rates, where the suprema runs over data distributions. In the special case of an interpolating classifier achieving zero empirical risk, the risk is shown to equal a certain mutual information term and to be controlled—for polynomial or slower rates—by upper bounds obtained by conditioning [**haghifam2022isit**].

Despite these successful applications, much less is known about the optimality or limitations of these techniques beyond the setting of binary classification and 0–1 valued loss. In this work, we turn our attention to stochastic convex optimization (SCO), a well-studied setting with known minimax rates, and look in particular at the analysis of stochastic gradient methods like stochastic gradient descent (SGD). In contrast to learning with a 0–1 valued loss, the minimax excess risk cannot be characterized in terms of uniform convergence of the generalization error [**shalev2010learnability**].

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

Next, we investigate whether the gap arises due to GD's deterministic nature: *can we close the gap by introducing randomness?* In other words, can we find a *"surrogate" algorithm* with good information-theoretic generalization guarantees, and such that this surrogate algorithm is close in generalization to the original one? Such an approach was formalized in [**negrea2020defense**; **pmlr-v178-sefidgaran22a**; **bu2021population**], and appears frequently in the generalization literature, e.g. [**neu2021information**; **wang2021generalization**; **harutyunyan2021information**; **LanCar2002**; **hellstrom2020nonvacuous**; **GR17**; **dziugaite2021role**; **neyshabur2018pac**; **zhang2020spread**; **chatterji2019intriguing**; **foret2020sharpness**; **wu2020adversarial**; **pour2022benefits**]. The most commonly-used surrogate is a "Gaussian surrogate", which perturbs the output of the algorithm by adding a Gaussian random variable. Surprisingly, we show that the limitations of information-theoretic analyses in the SCO setting are not eliminated even under the Gaussian surrogate.

Our negative results for Gaussian surrogates cast some doubt on their use to study SGD in other settings, such as deep learning. Information-theoretic techniques have shown some promise in this setting. Building off the seminal work of **PensiaJogLoh2018**, information-theoretic generalization bounds were shown to yield numerically nonvacuous estimates for stochastic gradient Langevin dynamics (SGLD) when applied to optimizing overparametrized neural networks on nontrivial deep learning classification benchmarks [**negrea2019information**; **haghifam2020sharpened**; **li2019generalization**; **pmlr-v162-banerjee22a**]. And while existing information-theoretic techniques seemingly cannot be applied directly to stochastic gradient methods like SGD itself, **neu2021information** showed how to obtain a (suboptimal) generalization bound for SGD using an information-theoretic bound for a noisy "surrogate" learning algorithm, designed to track the behavior of SGD. Our results explain the suboptimality of this approach and motivate work understanding the power or limitations of other surrogates.

1.1. Contributions

1. We prove *tight* generalization bounds based on the input-output mutual information (IOMI) of **RussoZou16** and **XuRaginsky2017** and the conditional mutual information (CMI) of **steinke2020reasoning** for CLB subclasses of SCO problems, as well as their individual sample variations [**negrea2019information**; **haghifam2020sharpened**; **bu2020tightening**; **rodriguez2020random**; **zhou2020individually**] and evaluated CMI [**steinke2020reasoning**]. Our generalization bounds may be of independent interest and can be used to obtain distribution- and algorithm-dependent generalization bounds for SCO problems beyond the worst-case guarantees.

2. We investigate whether we can *directly* analyze the generalization of GD with our information-theoretic generalization bounds. We provide a negative answer to this question by showing that neither the CMI nor IOMI frameworks can properly characterize the excess risk of GD in SCO problems in the minmax setting. We also extend our negative results to the alternative variations of IOMI and CMI, such as evaluated CMI [**steinke2020reasoning**], and individual sample bounds [**negrea2019information**; **haghifam2020sharpened**; **bu2020tightening**; **rodriguez2020random**; **zhou2020individually**].

3. We consider a surrogate algorithm based on a Gaussian perturbation of the final iterate of GD. We show that the generalization of GD can be decomposed as the sum of the generalization of the perturbed final iterate and a residual term that captures the sensitivity of the loss function to perturbations around such iterate. We consider a favorable setting where the parameters of the surrogate can be tuned based on the data distribution. Nevertheless, we show that there exists a sequence of CLB problems that can be learned with GD but IOMI and CMI frameworks fail to capture learnability in the minimax sense. Our construction is inspired from the ideas by **amir2021sgd** but with a completely different analysis.

4. We complement our results by showing that our construction also implies the failure of high-probability PAC-Bayes bounds in characterizing learnability of the CLB subclass of SCO problems using GD in the minimax sense. In particular, we prove that the *classical* PAC-Bayes bound of **McAllester1999** and the recently proposed *conditional* PAC-Bayes bound of **grunwald2021pac** are *vacuous* in the minimax sense.

1.2. Related Work

Recently, there has been a significant interest in understanding whether information-theoretic generalization bounds can characterize worst-case (minimax) rates for certain learning problems. For binary classification, **bassily2018learners**; **roishay** show that the IOMI and classical PAC-Bayes frameworks of [**RussoZou16**; **XuRaginsky2017**; **McAllester1999**] provably fail to characterize the learnability of Vapnik–Chervonenkis classes for which we have strong generalization guarantees. Then, **steinke2020reasoning**; **grunwald2021pac**; **haghifam2021towards** show that CMI [**steinke2020reasoning**] can be used to establish optimal bounds in the realizable setting. The results of [**steinke2020reasoning**; **grunwald2021pac**; **haghifam2021towards**] show that existing IT bounds *characterize* the minimax rates, *without the need for surrogates*. See also [**harutyunyan2022formal**; **9834457**; **nokleby2021information**]. Our work is different from the prior work since we study limitations of information-theoretic generalization bounds in the context of gradient descent methods. Moreover, our results indicate that existing techniques fail to characterize the minimax rates for gradient descent methods in SCO problems. Our findings stand in stark contrast to the success of information-theoretic frameworks in capturing the learnability of VC classes.

2. Preliminaries

2.1. Probability and Information Theory Notation

Let P, Q be probability measures on a measurable space. For a P-integrable function f, let $P[f] = \int f \mathrm{d}P$. 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 $\mathrm{KL}(Q\|P)$, is defined as $Q[\log \frac{\mathrm{d}Q}{\mathrm{d}P}]$ when $Q \ll P$ and as infinity otherwise.

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

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

Let $\mu = \mathbb{P}[X]$ and let $\kappa(Y) = \mathbb{P}^Y(X)$ a.s. If X concentrates on a countable set V with counting measure ν, the *(Shannon) entropy of X* is $\mathrm{H}(X) = -\mu[\log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}] = -\sum_{x\in V} \mathbb{P}(X = x) \log \mathbb{P}(X = x)$. The *disintegrated entropy of X given Y* is defined by $\mathrm{H}^Y(X) = -\kappa(Y)[\log \frac{\mathrm{d}\kappa(Y)}{\mathrm{d}\nu}]$, while the *conditional entropy of X given Y* is $\mathrm{H}(X|Y) = \mathbb{E}[\mathrm{H}^Y(X)]$. Note that $\mathrm{H}(X|Y) \leq \mathrm{H}(X)$ [**cover2012elements**].

2.2. Stochastic Convex Optimization

A *stochastic convex optimization* (SCO) problem is a triple $(\mathcal{W}, \mathcal{Z}, f)$, where $\mathcal{W} \subseteq \mathbb{R}^d$ is a convex set and $f(\cdot, z) : \mathcal{W} \to \mathbb{R}$ is a convex function for every $z \in \mathcal{Z}$ [**shalev2009stochastic**]. Informally, given an SCO problem $(\mathcal{W}, \mathcal{Z}, f)$, the goal is to find an approximate minimizer of the *population risk*

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

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

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

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

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

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

where the infimum runs over algorithms.[1]

In general, the class $\mathcal{L}$ itself is not learnable [**shalev2014understanding**]. One important family of subclasses of $\mathcal{L}$ which are known to be learnable are the convex–Lipschitz–bounded (CLB) subclasses of SCO problems where, for constants $L, R \in (0, \infty)$, the loss function $f(\cdot, z)$ is L-Lipschitz for all data instances $z \in \mathcal{Z}$, and the domain $\mathcal{W}$ is closed and has finite diameter R [**shalev2014understanding**]. We denote each such class of SCO problems by $\mathcal{C}_{L,R}$. In the remainder of the paper, we assume, without loss of generality, that each such $\mathcal{W}$ satisfies $\mathcal{W} \subseteq \{w : \|w\|_2 \leq R\}$.

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

$$\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) + \mathbb{E}\big[\hat{\mathrm{F}}_{S_n}(\mathcal{A}_n(S_n)) - \hat{\mathrm{F}}_{S_n}(W^\star_{S_n})\big] + \mathbb{E}\big[\hat{\mathrm{F}}_{S_n}(W^\star_{S_n}) - \mathrm{F}_{\mathcal{D}}(w^\star)\big],$$

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

The third term satisfies $\mathbb{E}[\hat{\mathrm{F}}_{S_n}(W^\star_{S_n}) - \mathrm{F}_{\mathcal{D}}(w^\star)] = \mathbb{E}[\hat{\mathrm{F}}_{S_n}(W^\star_{S_n}) - \hat{\mathrm{F}}_{S_n}(w^\star)] \leq 0$ because $W^\star_{S_n}$ is an ERM for the training set S_n, and $w^\star$ is a constant. Thus, it often suffices to characterize the expected generalization error and optimization error to obtain tight control of the excess risk. For approaches based on iterative optimization, the optimization error can, in many cases, be bounded by a convergence analysis [**bubeck2015convex**]. Therefore, the problem of controlling expected excess risk frequently amounts to controlling the expected generalization error. Nonetheless, there exist scenarios where the excess risk can vanish while the optimization and generalization errors do not, as shown in [**koren2022benign**] for some CLB problems learned with stochastic gradient descent (SGD).

CLB subclasses can be generically learned by suitably tuned instances of (projected) gradient descent (GD), a long studied algorithm [**cauchy1847methode**; **bubeck2015convex**]: For a convex and compact subset $\mathcal{W} \subseteq \mathbb{R}^d$, let $\Pi_{\mathcal{W}} : \mathbb{R}^d \to \mathcal{W}$ denote the Euclidean projection operator, given by $\Pi_{\mathcal{W}}(x) = \arg\min_{y\in\mathcal{W}} \|y - x\|_2$. The GD algorithm, $\mathrm{GD} = (\mathrm{GD}_n)_{n\geq 1}$, is initialized at some feasible point $W_0 \in \mathcal{W}$ and then, for some number T of iterations, proceeds to update the parameters iteratively according to $W_{t+1} = \Pi_{\mathcal{W}}\big(W_t - \eta_t g_t\big)$, where η_t is a suitably chosen

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

step-size and $g_t \in \partial \hat{\mathrm{F}}_{S_n}(W_t)$ is an element of the subdifferential of $\hat{\mathrm{F}}_{S_n}(W_t)$. While there are several variants, we will focus on the case where the output of the algorithm is the final iterate, i.e., $\mathrm{GD}_n(S_n) = W_T$.

2.3. Excess Risk of Gradient Descent

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

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

$$\sup_{(\mathcal{W},\mathcal{Z},f)\in\mathcal{C}_{L,R}} \sup_{\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})} \mathbb{E}\big[\hat{\mathrm{F}}_{S_n}(\mathrm{GD}_n(S_n)) - \hat{\mathrm{F}}_{S_n}(W^\star_{S_n})\big] \leq \frac{R^2}{2\eta T} + \frac{(\log(T)+2)\eta L^2}{2}. \tag{1}$$

(See Lemma 26 for a re-statement of this result in the context of the present paper). A similar result also appears in [**zhang2004solving**]. Recently, **bassily2020stability** proved a generalization bound for GD,

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

Together, Equations (1) and (2) yield the following bound on the excess risk,

$$\sup_{(\mathcal{W},\mathcal{Z},f)\in\mathcal{C}_{L,R}} \sup_{\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})} \mathbb{E}[\mathrm{F}_{\mathcal{D}}(\mathrm{GD}_n(S_n)) - \mathrm{F}_{\mathcal{D}}(w^\star)] \leq 4L^2\eta\left(\sqrt{T} + \frac{T}{n}\right) + \frac{R^2}{2\eta T} + \frac{(\log(T)+2)\eta L^2}{2}. \tag{3}$$

For all $\alpha \geq 2$, Equation (3) guarantees that GD achieves an excess risk in $\mathcal{O}(^{LR}/_{\sqrt{n}})$ for a number of iterations $T \in \Theta(n^\alpha)$ and a step-size $\eta \in \Theta(^{R\sqrt{n}}/_{Ln^\alpha})$. This, in fact, is the best achievable excess risk rate for the class $\mathcal{C}_{L,R}$ in the distribution-free setting [**bubeck2015convex**]. In [**amir2021sgd**; **sekhari2021sgd**], it is shown that GD cannot attain this excess risk rate when the number of iterations satisfies $T \in o\big(n^2\big)$.

3. Main Questions and Overview of the Results

The generalization error guarantee for GD in Eq. (2) is obtained using the algorithmic (uniform) stability framework of **bousquet2002stability**. (Prior work [**HardtRechtSinger2016**] also relied on algorithmic stability.) As shown above, a particular choice of the GD hyperparameters yields expected generalization error in $\mathcal{O}(^{LR}/_{\sqrt{n}})$. In this paper, we want to understand whether the same rate can be achieved using an information-theoretic framework for generalization. *Are information-theoretic frameworks for generalization expressive enough to accurately estimate the generalization error of GD for SCO?*

We begin by focusing on two frameworks for measuring the information complexity of a learning algorithm: *input-output mutual information* (IOMI [**XuRaginsky2017**; **RussoZou15**; **RussoZou16**]) and *conditional mutual information* (CMI [**steinke2020reasoning**]). The IOMI of an algorithm $\mathcal{A}_n$ with respect to a data distribution $\mathcal{D}$, denoted $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$, is defined to be the mutual information $I(\mathcal{A}_n(S_n); S_n)$ between the training data S_n and the output of the algorithm, $\mathcal{A}_n(S_n)$. In order to define the CMI framework, consider $n \in \mathbb{N}_+$ training data, let

$U = (U_1, \dots, U_n) \sim \mathrm{Unif}(\{0,1\}^n)$, and let $\tilde{S} = (\tilde{Z}_{i,j})_{i \in \{0,1\}, j \in \{n\}} \sim \mathcal{D}^{\otimes(2 \times n)}$ be a $2 \times n$ array of i.i.d. random elements in $\mathcal{Z}$, independent from U. Then $\tilde{S}_U = (\tilde{Z}_{U_i,i})_{i=1}^n$ has the same distribution as S_n, and so we may assume, w.l.o.g., that $S_n = \tilde{S}_U$ a.s. The CMI of the algorithm $\mathcal{A}_n$ with respect to the data distribution $\mathcal{D}$, denoted $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$, is defined to be the conditional mutual information $I(\mathcal{A}_n(S_n); U|\tilde{S})$ between $\mathcal{A}_n(S_n)$ and U given $\tilde{S}$. In the remainder of the section, we write $\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n)$ to refer to both $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$.

As the first step towards answering our main question, we develop new generalization bounds in both the IOMI and CMI frameworks to handle the CLB subclass of SCO, and show that our upper bounds are tight. Existing information-theoretic generalization bounds often depend on properties of the loss function $f(w,z)$ for fixed $w \in \mathcal{W}$. For instance, the generalization bounds in [**XuRaginsky2017**; **BuZouVeeravalli2019**; **negrea2019information**] depend on the tail of the random variable $f(w,Z)$ when $Z \sim \mathcal{D}$. In SCO, we often have no such control, making it impossible to reason about these problems using existing generalization bounds. Instead, in SCO, it is common for loss functions $f(w,z)$ to have regularity for fixed $z \in \mathcal{Z}$. In Theorem 2, we develop new information-theoretic generalization bounds for the $\mathcal{C}_{L,R}$ subclass, proving that $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathcal{O}\Big(LR\sqrt{\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n)/n}\Big)$. In Theorem 3, we show that our bound is *tight* up to constants.

Having obtained $\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n)$ bounds for SCO problems, we ask whether they capture the generalization properties of GD well enough to obtain minimax rates. In Section 5, we provide a negative answer to this question, proving that for sufficiently large n

$$\sup_{(\mathcal{W},\mathcal{Z},f) \in \mathcal{C}_{L,R}} \sup_{\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})} \mathrm{IM}_{\mathcal{D}}(\mathrm{GD}_n) \in \Omega(n), \tag{4}$$

which implies that neither the CMI nor IOMI frameworks can properly characterize minimax excess risk of GD in SCO problems. In Section 7, we study variations of IOMI and CMI, such as evaluated CMI [**steinke2020reasoning**] and individual-sample bounds [**bu2020tightening**; **negrea2019information**; **haghifam2020sharpened**; **rodriguez2020random**; **zhou2020individually**]. We find that they also fail to characterize the generalization of GD algorithm.

Since a direct analysis of GD via $\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n)$ is not possible, we consider a "surrogate" analysis [**negrea2020defense**], where an excess risk bound is obtained by comparing the risk of GD to a different (surrogate) algorithm, for which one can obtain generalization guarantees. In other words, *is GD "close" to an algorithm with small information complexity?*

We consider commonly used surrogate, whereby one perturbs the final iterate by a Gaussian random variable (see, e.g., [**hellstrom2020nonvacuous**; **GR17**; **wang2021generalization**; **neu2021information**; **dziugaite2021role**; **neyshabur2018pac**]). More formally, let $\mathcal{A}_n(S_n) = \tilde{W}$, where $\tilde{W} = \Pi_{\mathcal{W}}(W_T + \xi)$, $W_T = \mathrm{GD}_n(S_n)$, $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$, and $\xi \perp\!\!\!\perp S_n$. The generalization of GD can be related to the generalization of the Gaussian surrogate $\mathcal{A}_n$ via the inequality

$$\mathrm{EGE}_{\mathcal{D}}(\mathrm{GD}_n) \leq \mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) + \mathbb{E}[\Delta_\sigma(W_T)] + \mathbb{E}[\hat{\Delta}_\sigma(W_T)], \quad \text{where} \tag{5}$$

$$\Delta_\sigma(W_T) = \mathbb{E}^{S_n}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big] \text{ and } \hat{\Delta}_\sigma(W_T) = \mathbb{E}^{S_n}\Big[|\hat{\mathrm{F}}_{S_n}(\tilde{W}) - \hat{\mathrm{F}}_{S_n}(W_T)|\Big] \tag{6}$$

are referred to as *residual terms* in the sequel. In Eq. (6), the conditional expectations (given S_n) marginalize over only the (independent) randomness of ξ. Intuitively, the residual terms measure the sensitivity of the population and empirical loss landscapes [**neu2021information**]. The sensitivity is measured around W_T to perturbations by an isotropic Gaussian random vector with variance σ^2.

Remark 1 *In Eq. (5), one can drop the absolute values from the residual terms, i.e, second and third terms, to obtain*

$$\mathrm{EGE}_{\mathcal{D}}(\mathrm{GD}_n) = \mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) + \mathbb{E}\Big[\mathrm{F}_{\mathcal{D}}(W_T) - \mathrm{F}_{\mathcal{D}}(\tilde{W})\Big] + \mathbb{E}\Big[\hat{\mathrm{F}}_{S_n}(\tilde{W}) - \hat{\mathrm{F}}_{S_n}(W_T)\Big]. \tag{7}$$

In this remark, we want to demonstrate how tautologies can arise if one directly studies Eq. (7) instead of Eq. (5). Consider a surrogate that simply outputs a fixed parameter from $\mathcal{W}$ (independent of the training set). For instance, let $\mathcal{A}_n(S_n) = 0$ ($\tilde{W} = 0$). Then $\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n) = 0$ and $\hat{\mathrm{F}}_{S_n}(\tilde{W}) = \mathrm{F}_{\mathcal{D}}(\tilde{W})$. Therefore, Eq. (7) in this case is simplified to

$$\mathrm{EGE}_{\mathcal{D}}(\mathrm{GD}_n) = \mathbb{E}[\mathrm{F}_{\mathcal{D}}(W_T)] - \mathbb{E}\Big[\hat{\mathrm{F}}_{S_n}(W_T)\Big] = \mathrm{EGE}_{\mathcal{D}}(\mathrm{GD}_n), \tag{8}$$

taking us back to the original problem.

*Next, we argue that even if we restrict the surrogate algorithm to the case of perturbation by Gaussian random variable, i.e., $\mathcal{A}_n(S_n) = \tilde{W}$ where $\tilde{W} = \Pi_{\mathcal{W}}(W_T + \xi)$, we get an equally tautological statement from the decomposition in Eq. (7). In particular, we claim that letting $\sigma \to \infty$ takes us back to the original problem. Consider the IOMI framework. Since $\mathcal{W}$ is bounded, we have Var$(W_T) \in \mathcal{O}(1)$. Then, using [**polyanskiy2014lecture**] and the data-processing inequality for mutual information, we obtain $I(\Pi_{\mathcal{W}}(W_T + \xi); S_n) \leq I(W_T + \xi; S_n) \leq I(W_T + \xi; W_T) \in \mathcal{O}\big(1/\sigma^2\big)$ which tend to 0 as σ diverges. Also, as $\sigma \to \infty$, $\mathbb{E}[\hat{\mathrm{F}}_{S_n}(\tilde{W})] \approx \mathbb{E}[\mathrm{F}_{\mathcal{D}}(\tilde{W})]$. Therefore, in the case that $\sigma \to \infty$, by simplifying Eq. (7), we arrive at the same tautology as in Eq. (8). Since $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ for any learning problem [**haghifam2020sharpened**], we have the same tautology even if we use the CMI framework.*

In this Gaussian surrogate setting, the question of whether $\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n)$ bounds characterize $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ is equivalent to asking whether

$$\sup_{(\mathcal{W},\mathcal{Z},f)\in\mathcal{C}_{L,R}} \sup_{\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})} \inf_{\sigma\geq 0} \left\{ LR\sqrt{\frac{\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n)}{n}} + \mathbb{E}[\Delta_\sigma(W_T)] + \mathbb{E}[\hat{\Delta}_\sigma(W_T)] \right\} \overset{?}{\in} \Theta\Big(\frac{LR}{\sqrt{n}}\Big). \tag{9}$$

Answering this amounts to answering whether one can choose a value of σ *with a full knowledge of the SCO problem and the data distribution*, such that the perturbed GD algorithm achieves the optimal rate via the generalization bound appearing in Theorem 2. Alternatively, one can ask whether one can show, using the perturbation idea, that GD learns the subclass $\mathcal{C}_{L,R}$ *even with an arbitrary slow rate*, i.e., whether or not the LHS of Eq. (9) converges to zero as the number of the training samples diverges.

In order to gain insight on the Gaussian surrogate, consider extreme values of the variance of the perturbations. Setting $\sigma = 0$ corresponds to a direct analysis of GD, and the result in Eq. (4) shows we cannot prove learnability using existing frameworks. At the other extreme, one can show that $\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n) \to 0$ as $\sigma \to \infty$, leaving us with a bound in terms of the sum of the residual terms alone. As the distance between $\tilde{W}$ and W_T is maximal under such a perturbation, the sum of the residual terms is in $\Omega(1)$, once again failing to establish learnability. The idea behind introducing the surrogate algorithm $\mathcal{A}$ and adjusting the value of σ is that it allows one to conceptually interpolate between these two extreme points in order to find an optimal bound on GD's generalization error.

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

$$\sup_{(\mathcal{W},\mathcal{Z},f)\in\mathcal{C}_{L,R}}\ \sup_{\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})}\ \inf_{\sigma\geq 0}\left\{LR\sqrt{\frac{\mathrm{IM}_{\mathcal{D}}(\mathcal{A}_n)}{n}}+\mathbb{E}[\Delta_\sigma(W_T)]+\mathbb{E}[\hat{\Delta}_\sigma(W_T)]\right\}\in\Omega(1). \tag{10}$$

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

In Section 6, we extend our results to PAC-Bayes bounds, which provide tail bounds on the generalization error of GD, with respect to the randomness in the data. A similar surrogate decomposition as in Eq. (5) relates *disintegrated* generalization of GD to the generalization of $\mathcal{A}_n$ via

$$\mathbb{E}^{S_n}\left[\mathrm{F}_{\mathcal{D}}(W_T)-\hat{\mathrm{F}}_{S_n}(W_T)\right]\leq\mathbb{E}^{S_n}\left[\mathrm{F}_{\mathcal{D}}(\tilde{W})-\hat{\mathrm{F}}_{S_n}(\tilde{W})\right]+\hat{\Delta}_\sigma(W_T)+\Delta_\sigma(W_T), \tag{11}$$

where $\hat{\Delta}_\sigma(W_T)$ and $\Delta_\sigma(W_T)$ are defined in Eq. (6). The first term on the RHS of Eq. (11) can be analyzed using PAC-Bayes frameworks (see, e.g., [**LanCar2002**; **hellstrom2020nonvacuous**; **GR17**; **dziugaite2021role**; **neyshabur2018pac**; **chatterji2019intriguing**; **foret2020sharpness**; **wu2020adversarial**]). In Section 6, we show that, in the minimax sense, the classical and conditional PAC-Bayes frameworks of **McAllester1999** and **grunwald2021pac** provide a vacuous characterization of the RHS of Eq. (11) for all values of σ.

4. Information-Theoretic Generalization Bounds for the CLB setting

In SCO problems, generalization bounds for gradient methods can be obtained using the uniform stability framework [**HardtRechtSinger2016**; **bassily2020stability**; **feldman2018generalization**; **feldman2019high**]. This framework provides an algorithm-dependent approach that has been used to obtain relatively strong generalization bounds for several convex optimization algorithms in the distribution-free setting. In this section, we extend the CMI and IOMI frameworks to the CLB setting and provide *algorithm*- and *distribution*- dependent generalization bounds.

Theorem 2 *Let $n\in\mathbb{N}$, $\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})$ be a data distribution, and $S_n\sim\mathcal{D}^{\otimes n}$. Consider an SCO problem $(f,\mathcal{W},\mathcal{Z})\in\mathcal{C}_{L,R}$. Then, for every learning algorithm $\mathcal{A}_n$ such that $\mathcal{A}_n(S_n)\in\mathcal{W}$ a.s.,*

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

The proof, based on [**rodriguez2021tighter**], is in Appendix A. To better contextualize our generalization bounds in Theorem 2, we study their tightness. For the trivial case where the output of a learning algorithm is independent of the training set, the bounds in Theorem 2 are tight. The theorem below states that the bounds are tight even when the learning algorithm depends on the training set.

Theorem 3 *For every $n\in\mathbb{N}, L\in\mathbb{R}_+, R\in\mathbb{R}_+$, there exists an SCO problem $(f,\mathcal{W},\mathcal{Z})\in\mathcal{C}_{L,R}$, a data distribution $\mathcal{D}$ over $\mathcal{Z}$, and a learning algorithm $\mathcal{A}=(\mathcal{A}_n)_{n\geq 1}\in\mathcal{W}$ such that: (i) the expected generalization error of $\mathcal{A}_n$ satisfies $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)\geq {}^{LR}/_{\sqrt{2n}}$, and (ii) the upper bounds from Theorem 2 are $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)\leq {}^{LR\sqrt{2}}/_{\sqrt{n}}$ and $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)\leq {}^{LR\sqrt{8}}/_{\sqrt{n}}$, respectively.*

See Appendix B for the proof, which is inspired by [**orabona2019modern**]. Theorem 3 shows that there exists a learning algorithm in the CLB setting for which the bounds Theorem 2 is tight. This implies that the bound in Theorem 2 cannot be *uniformly* improved for every learning algorithm in the CLB setting. Note, however, that there may exist a tighter bound for some learning algorithms.

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

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

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

Theorem 4 *Let $n \in \mathbb{N}$, $T_{(n)} = 2n^2$, $\eta_{(n)} = \frac{1}{n\sqrt{5n}}$, and $\mathrm{d}_{(n)} = 3T2^n/4$. Then, there exists a universal constant $N^\star \in \mathbb{N}$ such that for every $n \geq N^\star$, there exist a sequence of SCO problems $\{(f_{(n)}, \mathcal{W}_{(n)}, \mathcal{Z}_{(n)}) \in \mathcal{C}_{4,1}\}_{n\in\mathbb{N}}$ where $\mathcal{W}_{(n)}, \mathcal{Z}_{(n)} \in \mathbb{R}^{d_{(n)}}$, and a data distribution $\mathcal{D}_{(n)}$ over $\mathcal{Z}_{(n)}$ such that the following holds: For $S_n \sim \mathcal{D}_{(n)}^{\otimes n}$, let $W_T = GD(S_n, \eta_{(n)}, T_{(n)})$ and $\mathcal{A}_n(S_n) = \Pi_{\mathcal{W}}(W_T + \xi)$, where $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_{(n)}})$. Then, there exists $\mathsf{var}^\star_n > 0$ such that if $\sigma^2 \leq \mathsf{var}^\star_{(n)}$, then $\mathrm{IOMI}_{\mathcal{D}_{(n)}}(\mathcal{A}_n) \in \Omega(n^3)$ and $\mathrm{CMI}_{\mathcal{D}_{(n)}}(\mathcal{A}_n) \in \Omega(n)$. Also, if $\sigma^2 > \mathsf{var}^\star_{(n)}$, then $\mathbb{E}[\hat{\Delta}_\sigma(W_T)] + \mathbb{E}[\Delta_\sigma(W_T)] \in \Omega(1)$. As a result,*

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

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

Proof Here, we provide an overview of the proof. The formal proof can be found in Appendix C. Our construction is inspired from the construction in **amir2021sgd**.

- **Construction and Dynamics of GD**: Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{0,1\}^d$. Let the data distribution on input be $(\mathrm{Ber}(1/2))^{\otimes d}$, i.e., each coordinate is drawn independently and uniformly at random from $\mathrm{Ber}(1/2)$. Thus, the training set $S_n \in \{0,1\}^{n\times d}$ is a matrix whose elements are drawn i.i.d. from $\mathrm{Ber}(1/2)$. Let λ be a sufficiently small constant, and $\mathcal{W}$ be a ball of radius one in $\mathbb{R}^d$. We consider the following loss function $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, $f(w,z) = \sum_{i=1}^d z(i)w(i)^2 + \lambda\langle w, z\rangle + \max\{\max_{i\in[d]}\{w(i)\}, 0\}$. We show that this function is convex and 4-Lipschitz. As a result, the problem is in the CLB subclass. Next, we demonstrate that when the dimension is $d = 3T2^n/4$, there are many columns in S_n such that *all* the entries are zero. Following **amir2021sgd**, we refer to such columns as *bad coordinates*. Let $\mathsf{B} \in \{0,1\}^d$ be a vector whose $i-$th coordinate is one if and only if i is a bad coordinate. We show that, with high probability, the number of bad coordinates is between $T/2$ and T. The result emerges from the observation that the dynamics of GD along the bad coordinates are completely *different* compared to the good coordinates, therefore "revealing" which coordinates are bad. To see this, consider the empirical risk $\hat{\mathrm{F}}_{S_n}(w) = \sum_{i=1}^d \hat{\mu}(i)w(i)^2 + \lambda\langle\hat{\mu}, w\rangle + \max\{\max_{i\in[d]}\{w(i)\}, 0\}$, where for

$i \in [d]$, $\hat{\mu}(i) = \frac{1}{n}\sum_{j=1}^{n} z_j(i) \in [0,1]$ is the empirical mean of the points in the i-th column of S_n. By the definition of the bad coordinates we can write $\hat{\mathrm{F}}_{S_n}(w) = \sum_{i\in\{i:\mathsf{B}[i]=0\}} \hat{\mu}(i)w(i)^2 + \lambda\sum_{i\in\{i:\mathsf{B}[i]=0\}} \hat{\mu}(i)w(i) + \max\{\max_{i\in[d]}\{w(i)\},0\}$. Note that the third term is not differentiable. We consider a specific first-order oracle proposed in [**amir2021sgd**; **bassily2020stability**]. We show that to analyze the dynamics of GD for good coordinates, we only need to consider the first two terms. For good coordinates, the main observation here is that because of the *norm-like* penalty from the first term, $|W_T(i)|$ is small. In contrast, for the bad coordinates the gradient that comes from the third term pushes $W_T(i)$ away from zero; in particular, for the bad coordinates we have $|W_T(i)| = \eta$ under the event $T/2 \le \|\mathsf{B}\|_0 \le T$. The other key property used in the proof is that $\|W_T\| \in \Theta(1/\sqrt{n})$ with high probability, meaning that the final iterate of GD is close to the origin.

- **Lower Bound on the Residual Term**: First, we prove that if $\sigma^2 \in \Omega(1/d)$, then the residual term is large. Recall that $\|W_T\| \in \Theta(1/\sqrt{n})$, and $\tilde{W} = \Pi_{\mathcal{W}}(W_T + \xi)$. Consider $\mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big]$, where the population risk is given by $\mathrm{F}_{\mathcal{D}}(w) = {}^1\!/_2\|w\|^2 + {}^\lambda\!/_2\sum_{i=1}^{d} w(i) + \max\{\max_{i\in[d]}\{w(i)\},0\}$. Using concentration inequalities for Gaussian random variables, we show that $\|\tilde{W}\|^2 \approx \min\{\sigma^2 d + o(1), 1\}$, while $\|W_T\|^2 \in o(1)$. Using this argument we show that unless $\sigma^2 \in \mathcal{O}(1/d)$, the residual term grows with n. Since d is exponentially large in n, we conclude that the variance of noise has to satisfy $\sigma^2 \in \mathcal{O}(2^{-n})$.

- **Lower Bound on $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ and $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$:** Here we show that $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \in \Omega(n)$, which implies that $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n) \in \Omega(n)$, since $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \le \mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ [**haghifam2020sharpened**]. In Appendix C, we prove the stronger result $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n) \in \Omega(n^3)$. Step one is to establish that $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \ge n - (\mathrm{H}(\mathsf{B}|\tilde{W},\tilde{S}) + \mathrm{H}(U|\tilde{W},\tilde{S},\mathsf{B})) \ge n - (\mathrm{H}(\mathsf{B}|\tilde{W}) + \mathrm{H}(U|\tilde{S},\mathsf{B}))$ using standard properties of mutual information. Next, we seek to upper bound $\mathrm{H}(\mathsf{B}|\tilde{W})$ and $\mathrm{H}(U|\tilde{S},\mathsf{B})$. We do so using *Fano's inequality* (Lemma 24) but in a way that differs from its conventional use. The intuition behind using Fano's inequality is as follows: if there exists an estimator that can be used to predict B using $\tilde{W}$, then the conditional entropy $\mathrm{H}(\mathsf{B}|\tilde{W}_T)$ is small. The same also holds for predicting U using $\tilde{W},\tilde{S},\mathsf{B}$. The core of the proof then rests on designing two estimators: one for estimating B using $\tilde{W}$, and another one for estimating U using $\tilde{S}$ and B. We construct explicit estimators for each, and demonstrate that their probability of error is small. Thus Fano's inequality implies that the entropy terms of interest are small. To construct the first estimator, we use two important properties: (i) the variance of noise satisfies $\sigma^2 \in \mathcal{O}(2^{-n})$, and (ii) for the good coordinates $|W_T(i)|$ is very small, while for the bad coordinates we have $|W_T(i)| \in \Theta\big(n^{-1.5}\big)$. The proposed estimator is based on comparing $|\tilde{W}(i)|$ with a threshold. We show that σ^2 is much smaller than $|W_T(i)|$ for the bad coordinates. As a result, the Gaussian noise does not *perturb* the bad coordinates significantly. Thus, the error probability of this estimator can be arbitrarily small as n diverges. For constructing the second estimator, remember that: (i) by definition, in each column of $\tilde{S}$ exactly one sample is chosen for the training set, and (ii) by the definition of the bad coordinates, we know that if $i \in [d]$ is a bad coordinate, then for all $Z \in S_n$, we have $Z(i) = 0$. Therefore, in every column of the supersample, either one or both of the samples have *zeros in all of the bad coordinates*. Our proposed estimator is as follows: whenever there is only one sample, the estimator can perfectly recover U for that column. In the case of two samples, the estimator makes a random guess. We show that the probability that there are two samples in a column such that both have zeros in all of the bad coordinates is $\Theta(2^{-n^2})$. Therefore, the estimator makes an error with small probability.

■

Remark 5 *The sequence of SCO problems that witnesses that lower bound for the IOMI and CMI frameworks is* the same. *Hence, a tight generalization bound cannot be achieved for every SCO problem by considering the best framework for that problem out of the IOMI and CMI frameworks.*

Remark 6 *Equation (3) provides a general result for the excess risk guarantee of GD for every number of iterations T and the step size η. GD obtains the excess risk and the generalization error guarantees of $\mathcal{O}\left(\frac{LR}{\sqrt{n}}\right)$ by setting $T \in \Theta(n^{\alpha})$ and $\eta \in \Theta\left(\frac{R\sqrt{n}}{Ln^{\alpha}}\right)$ for every $\alpha \geq 2$. In Theorem 4, we state the results only for $\alpha = 2$. However, the same construction can be used to prove the lower bounds in Theorem 4 for every $\alpha \geq 2$. This observation shows a stronger failure: for every parameter setting under which GD attains the optimal excess risk, the upper bound in Eq. (9) does not even converge to zero, i.e., $\Omega(1)$.*

Remark 7 *A notable property of the construction in Theorem 4 is that the Lipschitz constant of the loss function and the diameter of $\mathcal{W}$ do not grow with dimension. By a simple scaling, our result in Theorem 4 implies the lower bound stated in Eq. (10).*

6. Implications for PAC-Bayes Bounds

In this section, we show that our construction that witnesses the lower bounds in Theorem 4 reveals a limitation of PAC-Bayes bounds for learning SCO problems with GD. Using PAC-Bayes bounds to analyze the generalization of gradient methods via the surrogate algorithm that perturbs the final weight with a Gaussian random variable is a prevailing method in the literature [**LanCar2002**; **hellstrom2020nonvacuous**; **GR17**; **dziugaite2018data**; **dziugaite2021role**; **neyshabur2018pac**; **chatterji2019intriguing**; **foret2020sharpness**; **wu2020adversarial**]. This approach leads to non-vacuous estimates of the generalization gap for non-convex problems such as training modern deep learning models. However, we show that it fails for the CLB subclass of SCO problems.

We consider a classical PAC-Bayes bound [**mcallester1999some**] and a recently-proposed conditional PAC-Bayes bound [**grunwald2021pac**]. The main difference between the two is the *measure of complexity* that characterizes generalization. We can represent an algorithm $\mathcal{A}_n$ with a posterior distribution $Q : \mathcal{Z}^n \to \mathcal{M}_1(\mathcal{W})$. A complexity measure appearing in classical PAC-Bayes bounds is $C_{\text{clas}}(n) = \mathrm{KL}(Q(S_n) \| \mathbb{E}[Q(S_n)])$. The conditional PAC-Bayes bound relies on some additional structure. Let $S_n = (Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$ and $S'_n = (Z'_1, \dots, Z'_n) \sim \mathcal{D}^{\otimes n}$ such that $S_n \perp\!\!\!\perp S'_n$. For every $u = (u_1, \dots, u_n) \in \{0,1\}^n$, define $\tilde{S}_u = ((1-u_1)Z_1 + u_1 Z'_1, \dots, (1-u_n)Z_n + u_n Z'_n)$. The complexity measure for the conditional-PAC Bayes bound [**grunwald2021pac**] is $C_{\text{cond}}(n) = \mathbb{E}^{S_n}[\mathrm{KL}(Q(S_n) \| 2^{-n} \sum_{u \in \{0,1\}^n} Q(\tilde{S}_u))]$. Next we present the known results that relate these complexity measures to the generalization gap.

Theorem 8 ([mcallester1999some; grunwald2021pac]) *Let $S_n \sim \mathcal{D}^{\otimes n}$, $\delta \in (0,1)$, $L, R \in \mathbb{R}_+$. Assume that the range of the loss function f lies in $[-LR, LR]$. Then, with probability at least*

$(1-\delta)$ (over the choice of $S_n \sim \mathcal{D}^{\otimes n}$) for any posterior distribution $Q : \mathcal{Z}^n \to \mathcal{M}_1(\mathcal{W})$ with $W \sim Q(S_n)$,

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

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

Theorem 9 *Let $n \in \mathbb{N}$, $T_{(n)} = 2n^2$, $\eta_{(n)} = \frac{1}{n\sqrt{5n}}$, $d_{(n)} = 3T2^n/4$, and $N^\star \in \mathbb{N}$ be a universal constant. Then, there exists $\omega \in (0,1)$, a sequence of SCO problems $\{(f_{(n)}, \mathcal{W}_{(n)}, \mathcal{Z}_{(n)}) \in \mathcal{C}_{4,1}\}_{n\in\mathbb{N}}$ where $\mathcal{W}_{(n)}, \mathcal{Z}_{(n)} \in \mathbb{R}^{d_{(n)}}$, and a data distribution $\mathcal{D}_{(n)}$ over $\mathcal{Z}_{(n)}$ such that the following holds for all $n \geq N^\star$: For $S_n \sim \mathcal{D}_{(n)}^{\otimes n}$, let $W_T = GD(S_n, \eta_{(n)}, T_{(n)})$ and $\mathcal{A}_n(S_n) = \Pi_{\mathcal{W}}(W_T + \xi)$, where $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_{(n)}})$. Then, for every $0 < \delta < 1 - \omega$, with probability at least $1 - \delta - \omega$ over $S_n \sim \mathcal{D}_{(n)}^{\otimes n}$,*

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

This result implies that a PAC-Bayes bound for the surrogate from Eq. (11) yields a *vacuous* generalization bound with *constant probability*, i.e., independent of n.

Remark 10 *The complexity term in PAC-Bayes bounds generally takes the form $\mathrm{KL}(Q(S_n)\|P)$ for some element $P \in \mathcal{M}_1(\mathcal{W})$. The choice here, $P = \mathbb{E}[Q(S_n)]$, minimizes the complexity term in expectation. Whether other choices might yield tighter high probability bounds is left open.*

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

In the previous sections, we showed how the IOMI and the CMI frameworks and their high-probability counterparts fail to characterize the behavior of GD in the CLB setting, even when they are strengthened with a surrogate analysis. In this section, we consider other alternatives and reinforcements of these two frameworks and show that they also fail to characterize the behavior of SCO problems in the CLB setting, albeit without considering any potential strengthening with a surrogate analysis. First, we introduce these alternatives and their motivation, then we adapt them to the CLB setting, and finally we show their failure.

7.1. Information-Theoretic Alternatives to the IOMI and CMI Frameworks

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

1 The IOMI may be infinite and the CMI may be $\Omega(n)$ for a variety of learning scenarios, e.g., deterministic algorithms.

2 IOMI and CMI may capture unnecessary information. Note that we can write $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n) = \sum_{i=1}^{n} I(\mathcal{A}(S_n); Z_i) + I(Z^{i-1}, Z_i|\mathcal{A}(S_n))$, where $Z^{i-1} := (Z_1, \ldots, Z_{i-i})$. It is clear from this decomposition that IOMI not only captures the information that the output contains about individual samples, but also captures the “artificial" dependencies among the samples, given the algorithm's output. The latter is not predictive of the generalization performance of the algorithm [**bu2020tightening**]. An analogous problem arises in CMI, which includes the dependence of the indices and the samples after observing the algorithm's output [**rodriguez2020random**].

These problems can be avoided with an *individual-sample* bound proposed in [**bu2020tightening**], replacing $I(\mathcal{A}(S_n); S)$ with the average of $I(\mathcal{A}(S_n); Z_i)$ for all $i \in [n]$. This bound takes into account the information the output of the algorithm captures about each *individual* sample Z_i, disregarding the generated dependency between the samples after observing the said output. Similarly, the individual-sample bound from [**rodriguez2020random**; **zhou2020individually**] considers the information the output of the algorithm captures about each individual index U_i, disregarding the dependency between the indices and the samples. These bounds are adapted to the CLB setting in the following theorem. The proof is in Appendix A.

Theorem 11 *Let $n \in \mathbb{N}$, $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be a data distribution, and $S_n \sim \mathcal{D}^{\otimes n}$. Consider an SCO problem $(f, \mathcal{W}, \mathcal{Z}) \in \mathcal{C}_{L,R}$. Then, for every learning algorithm $\mathcal{A}_n$ such that $\mathcal{A}_n(S_n) \in \mathcal{W}$ a.s., we have $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq {}^{LR}\!/_{n} \sum_{i=1}^{n} \sqrt{2I(\mathcal{A}(S_n); Z_i)}$, and $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq {}^{2LR}\!/_{n} \sum_{i=1}^{n} \sqrt{2I(\mathcal{A}(S_n); U_i|\tilde{Z}_{0,i}, \tilde{Z}_{1,i})}$.*

Remark 12 *As mentioned above, the individual-sample alternatives to the IOMI and CMI are tighter than the IOMI and CMI. This may be seen by [**bu2020tightening**] and [**rodriguez2020random**] or [**zhou2020individually**], where we have that*

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

*Therefore, Theorem 11 implies Theorem 2. Moreover, we have that $I(\mathcal{A}(S_n); U_i|\tilde{Z}_{1,i}, \tilde{Z}_{2,i}) \leq I(\mathcal{A}(S_n), Z_i)$ [**rodriguez2021tighter**].*

Another drawback of IOMI and CMI frameworks is the following:

3 Both the IOMI and CMI of an algorithm depend on the joint distribution of the algorithm's output and other variables. In contrast, generalization error depends on the algorithm's output only through the losses it incurs. Therefore, it is possible to increase both the IOMI and the CMI by *embedding* information about the training set in the output of a learning algorithm without affecting the algorithm's statistical properties [**roishay**; **bassily2018learners**].

steinke2020reasoning propose an alternative framework, *evaluated* CMI, that considers the information about the data captured by the *incurred loss* rather than the output itself.

Definition 13 (Evaluated CMI, [steinke2020reasoning]) *Let $n \in \mathbb{N}$. Let the supersample $\tilde{S}$ and indices U be as defined in Section 3. Let $S_n = (Z_{U_i,i})_{i\in[n]}$, and $F \in \mathbb{R}^{2\times n}$ be the array with entries $F_{v,i} = f(\mathcal{A}_n(S_n), Z_{v,i})$ for $v \in \{0,1\}$, $i \in [n]$. The* evaluated conditional mutual information of $\mathcal{A}$ with respect to $\mathcal{D}$, *denoted by $e\mathrm{CMI}_{\mathcal{D}}(f(\mathcal{A}_n))$, is the conditional mutual information $I(F; U|\tilde{S})$.*

haghifam2021towards show that the eCMI can provide a sharp characterization of generalization for the realizable setting and 0–1 losses. Below, we state a bound for the CLB setting based on eCMI. The proof can be found in Appendix A.

Theorem 14 *Consider an SCO problem* $(f, \mathcal{W}, \mathcal{Z}) \in \mathcal{C}_{L,R}$*. Then, for every learning algorithm* $\mathcal{A}_n$ *such that* $\mathcal{A}_n(S_n) \in \mathcal{W}$ *a.s., we have* $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq LR\sqrt{\frac{8\mathrm{eCMI}_{\mathcal{D}}(f(\mathcal{A}_n))}{n}}$.

Remark 15 *Similarly to Remark 12, note that the evaluated version of the CMI is tighter than the CMI itself, i.e.,* $\mathrm{eCMI}_{\mathcal{D}}(f(\mathcal{A}_n)) \leq \mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ *[***steinke2020reasoning***].*

Remark 16 *Since these alternatives to the IOMI and CMI are tighter than the IOMI and CMI themselves (cf. Remark 12 and Remark 15), the adaptation of these bounds to the CLB setting (Theorems 11 and 14) are also tight in the sense of Theorem 3.*

7.1.1. DATA-DEPENDENT ALTERNATIVES AND FUNCTIONAL CMI

negrea2019information and **haghifam2020sharpened** introduced data-dependent alternatives to the IOMI and CMI frameworks that resulted in numerically non-vacuous generalization guarantees for stochastic gradient Langevin dynamics (SGLD) and its full-batch counterpart for modern deep-learning datasets and architecture. These bounds can also be adapted to the CLB setting by replicating Theorem 11 (i) considering [**rodriguez2021tighter**] instead of [**rodriguez2021tighter**] and noting that $\mathbb{E}\big[\mathrm{KL}(\mathbb{P}^{S_n}(W) \| \mathbb{P}^{S_n \setminus Z_i}(W))\big] = I(W; Z_i | S_n \setminus Z_i)$; and (ii) considering [**rodriguez2021tighter**] instead of [**rodriguez2021tighter**] and noting that $\mathbb{E}\big[\mathrm{KL}(\mathbb{P}^{\tilde{S},U}(W) \| \mathbb{P}^{\tilde{S},U \setminus U_i}(W))\big] = I(W; U_i | \tilde{S}, U \setminus U_i)$. This would yield the data-dependent EGE bounds

$$\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{LR}{n} \sum_{i=1}^{n} \sqrt{2I(\mathcal{A}(S_n); Z_i | S_n \setminus Z_i)} \quad \text{and} \tag{12}$$

$$\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{2LR}{n} \sum_{i=1}^{n} \sqrt{I(\mathcal{A}(S_n), U_i | \tilde{S}, U \setminus U_i)}. \tag{13}$$

Both Eq. (12) and Eq. (13) are looser than the bounds in Theorem 11, by similar arguments to those in Remark 12 [**rodriguez2020random**].

harutyunyan2021information introduced an alternative to the CMI for supervised learning problems that yield bounds that can be experimentally computed and are non-vacuous. However, by the data processing inequality we have that this notion is looser than the evaluated CMI.

7.2. Failure of the Alternatives

We demonstrate that the individual sample and evaluated versions of CMI still fail in the CLB setting. Based on the relative tightness of these alternative frameworks (see Remark 12 and Remark 15), showing their failure implies failure of all the aforementioned alternatives to the IOMI and CMI frameworks. In fact, it also proves the failure of (i) the data-dependent bounds from [**negrea2019information**] and [**haghifam2020sharpened**], and (ii) functional-CMI of [**harutyunyan2021information**], adapted to the CLB setting (c.f. Section 7.1.1).

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

Theorem 17 *Let $n \in \mathbb{N}$, $T_{(n)} = n^2$, $\eta_{(n)} = \frac{1}{n\sqrt{n}}$, and $d_{(n)} = 2n^2$. Then, for every $n \geq 1$, there exists a sequence of SCO problems $\{(f_{(n)}, \mathcal{W}_{(n)}, \mathcal{Z}_{(n)}) \in \mathcal{C}_{1,1}\}_{n\in\mathbb{N}}$ where $\mathcal{W}_{(n)}, \mathcal{Z}_{(n)} \in \mathbb{R}^{d_{(n)}}$, and data distribution $\mathcal{D}_{(n)}$ over $\mathcal{Z}_{(n)}$ such that the following holds: Let $W_T = GD(S_n, \eta_{(n)}, T_{(n)})$. Then, $e\mathrm{CMI}_{\mathcal{D}_{(n)}}(f(\mathcal{A}_n)) \in \Omega(n)$, and $\sum_{i=1}^{n} \sqrt{2I(\mathcal{A}(S_n); U_i | \tilde{Z}_{0,i}, \tilde{Z}_{1,i})} \in \Omega(n)$, while the generalization error of GD satisfies $\mathbb{E}[|\mathrm{F}_{\mathcal{D}_{(n)}}(W_T) - \hat{\mathrm{F}}_{S_n}(W_T)|] \in \mathcal{O}(1/\sqrt{n})$.*

Proof Here, we provide an overview of the proof. The formal proof can be found in Appendix E. Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{\mathsf{e}(i) : i \in [d]\}$, where $\mathsf{e}(i) = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 at the i-th coordinate and $\|\mathsf{e}(i)\|_2 = 1$. Let the data distribution on the input be the uniform distribution, that is $\mathcal{D} = \mathrm{Uniform}(\mathcal{Z})$. Consider a problem in the CLB class with a convex, 1-Lipschitz loss function $f(w, z) = -\langle w, z\rangle$, and $\mathcal{W} = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$. With this loss, the weights W_T returned by GD after T iterations are a weighted sum of the instances Z_i. As in the *birthday paradox* [**mitzenmacher2017probability**] problem, we can show that for large d, e.g. $d = 2n^2$, the probability that any two instances from the supersample $\tilde{S}$ share the same non-zero coordinate is smaller than some constant probability c, which is independent of the number of samples. Let E be an $\tilde{S}$-measurable random variable that is one if and only if no pair of instances $\tilde{Z}_{u,i}$ and $\tilde{Z}_{v,j}$ (for all $i, j \in [n]$ and all $u, v \in \{0, 1\}$) from the supersample $\tilde{S}$ share the same coordinate.

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

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

8. Open Questions

In this work, we uncover the limitations of information-theoretic analyses of GD for the CLB subclass of SCO problems. We further show that these limitations remain even when a surrogate algorithm based on Gaussian perturbation is considered. Our results prompt several directions for future research:

1. **Optimal dependence of the information-theoretic bounds on the dimension**: One of the common properties between our constructions in Theorem 4 and Theorem 17 is that the dimension is much larger than the number of samples. In particular, we exploit the fact that the generalization guarantees of GD for SCO problems is dimension-independent in order to construct problem instances with large information complexity. In particular, it is straightforward to see that the lower bounds on IOMI and CMI that stem from our constructions in Theorem 4 and Theorem 17 depend on the dimension. It is interesting to find the minimum

dimension such that there exists an SCO problem for which the information-theoretic bounds fail to characterize learnability. For the direct analysis of GD we show $\mathcal{O}\big(n^2\big)$ is sufficient (Theorem 17), while for the surrogate analysis exponential dependence, i.e., $\mathcal{O}\big(n^2 2^n\big)$ (Theorem 4), is sufficient where n is the number of training samples.

2. **Gaussian perturbation for other subclasses of SCO problems**: In this work, we proved limitations of the surrogate algorithm based on the Gaussian perturbation for the CLB subclass of SCO problems. In particular, the loss function used in Theorem 4 is a *non-smooth* convex function. It is an open question to show that such limitations exist for the subclasses of SCO problems with smooth or strongly-convex loss functions. Notice that our results in Theorem 17 suggests that a *direct* analysis still fails for the subclass of SCO problems with smooth loss functions as the loss function used for proving Theorem 17 is smooth.

3. **Instance-independent surrogates**: The notable property of Gaussian perturbation is that it is instance-independent, in the sense that its structure does not depend on the problem instance, and we only need to tune the variance based on the problem instance. It is an open question to prove or refute the existence of a *instance-independent surrogate* for analyzing the generalization of gradient descent methods for SCO problems using information-theoretic frameworks. An interesting starting point is investigating the prospect of using the Gibbs algorithm [**wang2016average**; **aminian2021exact**] as a problem-independent surrogate.

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

Acknowledgments

The authors would like to thank Gergely Neu and Thomas Steinke for helpful discussions.

Disclosure of funding

Mahdi Haghifam is supported in part by the Vector Institute and Doctoral Completion Award from University of Toronto. Borja Rodríguez-Gálvez and Mikael Skoglund were funded in part by the Swedish research council under contract 2019-03606. Ragnar Thobaben was funded in part by the Swedish research council under contract 2021-05266. Part of the work was done while Mahdi Haghifam was an intern at Google Research-Brain Team and Borja Rodríguez-Gálvez was an intern at Apple. Daniel Roy is supported in part by an NSERC Discovery Grant and Accelerator. This research was carried out, in part, while Daniel Roy and Gintare Karolina Dziugaite were visiting the Simons Institute for the Theory of Computing at UC Berkeley, and while Daniel Roy was visiting the Institute for Mathematical Research at ETH Zürich.

Appendix A. Proof of the information-theoretic bounds of $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ in the CLB setting

Before starting the proofs, note that the proof of Theorem 11 implies Theorem 2 (c.f. Remark 12 and Remark 15).

A.1. Proof of Theorem 11: Individual-sample IOMI

Consider [**rodriguez2021tighter**], which controls $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ by means of the Wasserstein distance

$$\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{L}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{W}(\mathbb{P}^{Z_i}(W), \mathbb{P}(W))\Big].$$

Then, consider the fact that the Wasserstein distance is dominated by the total variation, that is, that $\mathbb{W}(\mu,\nu) \leq 2R\mathsf{TV}(\mu,\nu)$ when the space where the distributions μ and ν are defined has diameter R with respect to the specified metric [**villani2009optimal**][2]. Applying Pinsker's [**polyanskiy2014lecture**] inequality to the total variation and Jensen's inequality afterwards, one recovers the desired bound in Theorem 11.

A.2. Proof of Theorem 11: Individual-sample CMI

Consider now [**rodriguez2021tighter**], which again controls $\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n)$ by means of the Wasserstein distance

$$\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) \leq \frac{L}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{W}(\mathbb{P}^{U_i,\tilde{Z}_{0,i},\tilde{Z}_{1,i}}(W), \mathbb{P}^{\tilde{Z}_{0,i},\tilde{Z}_{1,i}}(W))\Big].$$

As in the proof above, considering the domination of the Wasserstein distance by the total variation together with Pinsker's and Jensen's inequality recovers the desired bound in Theorem 11.

A.3. Proof of Theorem 14

By the Donsker-Varadhan lemma [**boucheron2013concentration**] we have that

$$I(F,\tilde{S};U) \geq \mathbb{E}[g(F,\tilde{S},U)] - \log\mathbb{E}\Big[e^{g(F',\tilde{S}',U)}\Big]$$

for all measurable functions g such that $g(F,\tilde{S},U)$ and $e^{g(F',\tilde{S}',U)}$ have finite expectations [**boucheron2013concentration**], where $(F',\tilde{S}')$ is an independent copy of $(F,\tilde{S})$ and where $I(F,\tilde{S};U) = I(F,U|\tilde{S}) = e\mathrm{CMI}_{\mathcal{D}}(f(\mathcal{A}_n))$. For the rest of the proof, let $\boldsymbol{f} \in \mathbb{R}^{2\times n}$ be a realization of F. Consider now

$$g(\boldsymbol{f},\tilde{s},u) = \frac{\lambda}{n}\sum_{i=1}^{n}(2u_i-1)\Big(\boldsymbol{f}_{0,i} - f(0,\tilde{z}_{0,i}) - \big(\boldsymbol{f}_{1,i} - f(0,\tilde{z}_{1,i})\big)\Big)$$

[2] In the particular case of this work, the metric considered for the Lipschitness of the function and the diameter of the space is the ℓ_2 norm difference, but these theorems are not restricted to that.

for some $\lambda > 0$, and note that $\mathbb{E}[g(F, \tilde{S}, U)] = \lambda \text{EGE}_{\mathcal{D}}(\mathcal{A}_n)$. Applying Donsker-Varadhan lemma [**boucheron2013concentration**] with this choice of g yields

$$e\text{CMI}_{\mathcal{D}}(f(\mathcal{A}_n)) \geq \lambda \text{EGE}_{\mathcal{D}}(\mathcal{A}_n) - \log \mathbb{E}\left[e^{\frac{\lambda}{n}\sum_{i=1}^{n}(2U_i-1)\left(F'_{0,i}-f(0,\tilde{Z}'_{0,i})-\left(F'_{1,i}-f(0,\tilde{Z}'_{1,i})\right)\right)}\right].$$

Studying random variables $(2U_i - 1)\Big(F'_{0,i} - f(0, \tilde{Z}'_{0,i}) - \big(F'_{1,i} - f(0, \tilde{Z}'_{1,i})\big)\Big)$ reveals that they are 0 mean and bounded in $[-2LR, 2LR]$. We can thus apply Hoeffding's lemma [**wainwright2019high**] to bound the cumulant generating function. Optimizing for $\lambda > 0$ and rearranging completes the proof.

Appendix B. Proof of Theorem 3

Let $d \in \mathbb{N}$ be arbitrary. Let $\mathcal{W}$ be a ball of radius R in $\mathbb{R}^d$. Consider an arbitrary $z_0 \in \mathcal{W}$ such that $\|z_0\| = R$. The input space is $\mathcal{Z} = \{z_0/R, -z_0/R\}$. Also, let the data distribution $\mathcal{D}$ be $\mathcal{D}(z_0/R) = \mathcal{D}(-z_0/R) = 1/2$, the loss function $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ be $f(w, z) = -L\langle w, z\rangle$. It is straightforward to see that the loss function is convex and L-Lipschitz [3].

Denote the training set $S_n = (Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}$. Define a Rademacher random variable $\epsilon_i = 1$ if $Z_i = z_0/R$ and $\epsilon_i = -1$ if $Z_i = -z_0/R$. We can represent the training set as $S_n = (\frac{z_0}{R}\epsilon_1, \ldots, \frac{z_0}{R}\epsilon_n)$. The empirical risk for $w \in \mathcal{W}$ is given by $\hat{\mathrm{F}}_{S_n}(w) = \dfrac{-L}{nR}\langle w, z_0 \sum_{i\in[n]} \epsilon_i\rangle$. It is straightforward to see that the ERM for this problem is

$$\operatorname*{arg\,min}_{w\in\mathcal{W}} \hat{\mathrm{F}}_{S_n}(w) = \mathcal{A}_n(S_n) = \begin{cases} z_0 & \text{if } \operatorname{sign}(\sum_{i=1}^{n} \epsilon_i) = 1 \\ -z_0 & \text{if } \operatorname{sign}(\sum_{i=1}^{n} \epsilon_i) = -1 \end{cases},$$

where for $x \in \mathbb{R}$, $\operatorname{sign}(x) = 1$ if $x \geq 0$ and $\operatorname{sign}(x) = -1$ if $x < 0$.

First, we provide a lower bound on the expected generalization error. The expected empirical risk of $\mathcal{A}_n$ is given by

$$\begin{aligned}
\mathbb{E}\big[\min_{w\in\mathcal{W}} \hat{\mathrm{F}}_{S_n}(w)\big] &= \mathbb{E}\Big[\min_{w\in\mathcal{W}} -\frac{L}{Rn}\Big\langle w, z_0 \sum_{i=1}^{n} \epsilon_i\Big\rangle\Big] \\
&= -\frac{L}{Rn}\mathbb{E}\Big[\max_{w\in\{z_0,-z_0\}} \Big\langle w, z_0 \sum_{i=1}^{n} \epsilon_i\Big\rangle\Big] \\
&= -\frac{L}{Rn}\mathbb{E}\Big[\Big|\Big\langle z_0, z_0 \sum_{i=1}^{n} \epsilon_i\Big\rangle\Big|\Big] \\
&= -\frac{LR}{n}\mathbb{E}\Big[\Big|\sum_{i=1}^{n} \epsilon_i\Big|\Big],
\end{aligned}$$

[3] The construction for this section is inspired by the lower bounds for online convex optimization in [**orabona2019modern**].

where we have used $\forall a, b \in \mathbb{R}$, $\max(a,b) = \frac{a+b}{2} + \frac{|a-b|}{2}$. Observe that $\mathrm{F}_{\mathcal{D}}(w) = 0$ for all $w \in \mathcal{W}$. Therefore, the expected generalization error is lower bounded by

$$\begin{aligned}\mathrm{EGE}_{\mathcal{D}}(\mathcal{A}_n) &= -\mathbb{E}\Big[\min_{w\in\mathcal{W}} \hat{\mathrm{F}}_{S_n}(w)\Big]\\ &= \frac{LR}{n}\mathbb{E}\Big[\Big|\sum_{i=1}^{n}\epsilon_i\Big|\Big]\\ &\geq \frac{LR}{\sqrt{2n}},\end{aligned}$$

where the last line follows from Khintchine–Kahane inequality [**mohri2018foundations**].

Next, we analyze the upper bounds based on Theorem 2. Observe that the following Markov chain holds:

$$S_n - \mathrm{sign}\Big(\sum_{i=1}^{n}\epsilon_i\Big) - \mathcal{A}_n(S_n).$$

By the data processing inequality we have

$$\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n) = I(\mathcal{A}_n(S_n); S_n) \leq I(\mathcal{A}_n(S_n); \mathrm{sign}\Big(\sum_{i=1}^{n}\epsilon_i\Big)).$$

We can upper bound the mutual information as

$$I(\mathcal{A}_n(S_n); \mathrm{sign}\Big(\sum_{i=1}^{n}\epsilon_i\Big)) \leq \mathrm{H}(\mathrm{sign}\Big(\sum_{i=1}^{n}\epsilon_i\Big)) \leq 1,$$

since $\mathrm{sign}\Big(\sum_{i=1}^{n}\epsilon_i\Big)$ can take only two values. Therefore, we obtain $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 1$. As $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq \mathrm{IOMI}_{\mathcal{D}}(\mathcal{A}_n)$ for any learning problem [**haghifam2020sharpened**], we have $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 1$. Finally, the result follows by plugging the bounds on IOMI and CMI into Theorem 2.

Appendix C. Proof of Theorem 4

The outline of the proof is as follows. First, in Appendix C.1, we describe our construction. Then, we analyze the dynamics of GD on the problem in Appendix C.2. Using the properties of the final iterate of GD, proved in Appendix C.2, we proceed by showing in Appendix C.3 that if the noise variance is greater than a threshold, then the residual term does not converge to zero as the number of samples grows. For the case that the noise variance is smaller than the threshold, we prove the failure of IOMI and CMI in Appendix C.4 and Appendix C.5, respectively.

C.1. Construction

We begin the proof by describing a learning scenario that witnesses the lower bound (we drop the n argument from the parameters to reduce notational clutter). Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{0,1\}^d$. Let the data distribution on input be $(\mathrm{Ber}(1/2))^{\otimes d}$, i.e., each coordinate is drawn independently and uniformly at random from $\mathrm{Ber}(1/2)$. In this section, we treat the training set $S_n \in \{0,1\}^{n\times d}$ as a matrix. Note that each element of S_n is drawn i.i.d. from $\mathrm{Ber}(1/2)$. For $i \in [d]$, we say the $i-$th coordinate is a *bad coordinate* iff for all $j \in [n]$, $Z_j(i) = 0$. In words, if $i-$th coordinate is a bad coordinate then all the entries in the $i-$th column of S_n is zero. Also, the convex domain space $\mathcal{W}$ is the Euclidean ball of radius one in $\mathbb{R}^d$. Note for $x \in \mathbb{R}^d$, $\Pi_{\mathcal{W}}(x) = x/\max\{\|x\|, 1\}$.

We consider the convex function proposed in **amir2021sgd**. Let $0 < \lambda \leq \mathcal{O}\Big(1/(n\sqrt{d})\Big)$ be a positive constant which is determined later. Then we consider the following loss function $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$

$$f(w, z) = \sum_{i=1}^{d} z(i)w(i)^2 + \lambda\langle w, z\rangle + \max\Big\{\max_{i\in[d]}\{w(i)\}, 0\Big\}. \tag{14}$$

It is straightforward to show that the first two terms in Eq. (14) is convex. Also, $\max\{\max_{i\in[d]}\{w(i)\}, 0\}$ is a convex function because it is maximum of convex (linear) functions [**boyd2004convex**]. Therefore, f is convex as it is sum of convex functions. Then, we show that each term in Eq. (14) is Lipschitz. The first term, $\sum_{i=1}^{d} z(i)w(i)^2 \leq \|w\|^2$ is $2-$Lipschitz by the boundedness of $\mathcal{W}$. The second term is $\lambda\sqrt{d}-$Lipschitz because $\|\nabla(\lambda\langle w, z\rangle)\| = \lambda\|z\| \leq \lambda\sqrt{d}$. We use Lemma 25 to show that the last term in Eq. (14) is 1-Lipschitz. Therefore, $f(w, z)$ is $(3 + \lambda\sqrt{d})-$Lipschitz. Note that $\lambda \in \mathcal{O}\Big(1/(n\sqrt{d})\Big)$, so the function in Eq. (14) is $4-$Lipschitz for sufficiently large n.

C.2. Dynamics of GD

First of all, we want to note that the statements in this proof about random variables hold almost surely. We will skip such declarations for the remainder of the proof to aid readability. In this part, we aim to find the properties of the final iterates of the GD algorithm. Let $d = 0.75T2^n$. Let $\mathsf{B} \in \{0, 1\}^d$ denote a vector such that $\mathsf{B}(i) = 1$ if and only if i is a bad coordinate. Let $\|\mathsf{B}\|_0$ denote the number of bad coordinates. Next, we provide a probabilistic estimate on $\|\mathsf{B}\|_0$. $\|\mathsf{B}\|_0 = \sum_{i=1}^{d} \mathsf{B}(i)$ follows the binomial distribution with the number of trial d and the success probability of 2^{-n}. The reason is the probability that all the points in a column is zero is given by 2^{-n}. By the standard multiplicative Chernoff bound [**mitzenmacher2017probability**] we have

$$\mathbb{P}(T/2 \leq \|\mathsf{B}\|_0 \leq T) \geq 1 - 2\exp(-T/36). \tag{15}$$

Therefore, with probability at least $1 - 2\exp(-T/36)$, the number of bad coordinates is between $T/2$ and T.

Next step concerns understanding the dynamics of GD. The empirical risk for any $w \in \mathcal{W}$ is given by

$$\hat{\mathrm{F}}_{S_n}(w) = \sum_{i=1}^{d} \hat{\mu}(i)w(i)^2 + \lambda\langle\hat{\mu}, w\rangle + \max\Big\{\max_{i\in[d]}\{w(i)\}, 0\Big\} \tag{16}$$

where for $i \in [d]$, $\hat{\mu}(i) = \frac{1}{n}\sum_{j=1}^{n} z_j(i) \in [0, 1]$ is the empirical mean of the points in $i-$th column of S_n.

Lemma 18 *Under the event* $\{T/2 \leq \|\mathsf{B}\|_0 \leq T\}$, *let* $\mathcal{B} = \{i_1, \ldots, i_{\|\mathsf{B}\|_0}\} \subseteq [d]$ *contain the ordered set of bad coordinates. Consider the GD process* $W_{t+1} = W_t - \Pi_{\mathcal{W}}(W_t - \eta\partial(\hat{\mathrm{F}}_{S_n}(W_t)))$

starting at $W_0 = 0$ where η is the step size. For every $i \in [d]$ and $t \in [T]$

$$W_t(i) = \begin{cases} \frac{\lambda}{2}(-1 + (1 - 2\eta\hat{\mu}(i))^t) & i \in [d] \setminus \mathcal{B} \\ -\eta & i \in \{i_1, \dots, i_{\min\{\|\mathsf{B}\|_0, t-1\}}\} \\ 0 & i \in \{i_{\min\{\|\mathsf{B}\|_0, t-1\}+1}, \dots, i_{\|\mathsf{B}\|_0}\} \end{cases}.$$

In particular,

$$W_T(i) = \begin{cases} \frac{\lambda}{2}(-1 + (1 - 2\eta\hat{\mu}(i))^T) & i \in [d] \setminus \mathcal{B} \\ -\eta & i \in \mathcal{B} \end{cases},$$

and for all $i \in [d] \setminus \mathcal{B}$, $-\eta\lambda T \leq W_T(i) < 0$.

Proof First, we describe the first-order oracle proposed in **amir2021sgd** and **bassily2020stability**. Note that the first two terms in Eq. (14) are differentiable. For the third term, i.e., $f_3(w) = \max\{\max_{i\in[d]}\{w(i)\}, 0\}$, which is not differentiable we consider the following first-order oracle. Let $\mathcal{I}(w) = \{j \in [d] | j \in \{\arg\max_{i\in[d]} w(i)\} \cap \{i | w(i) \geq 0\}\}$. Then, we claim that

$$\partial f_3(w) = \begin{cases} 0 & w = 0 \text{ or } \mathcal{I} = \emptyset \\ \mathsf{e}(\min\{\mathcal{I}(w)\}) & w \neq 0 \text{ and } \mathcal{I} \neq \emptyset. \end{cases} \tag{17}$$

where for $i \in [d]$, $\mathsf{e}(i) = (\underbrace{0, \dots, 0}_{i-1 \text{ times}}, 1, \underbrace{0, \dots, 0}_{d-i \text{ times}})$ (i-th coordinate vector).

To prove that Eq. (17) is a member of subgradient at w, we need to prove for all $w, v \in \mathbb{R}^d$, we have $f_3(v) \geq f_3(w) + \langle \partial f_3(w), v - w \rangle$.

Consider the case $w = 0$, then, since $\partial f_3(w) = 0$, trivially we have $f_3(v) \geq f_3(w) = 0$. Next, consider the case that $w \neq 0$ but $\mathcal{I}(w) = \emptyset$. This case holds if and only if for all $i \in [d]$, $w(i) < 0$. Therefore, $\partial f_3(w) = 0$ and the first-order convexity condition trivially holds. Finally, consider the case that $w \neq 0$ and $\mathcal{I}(w) \neq \emptyset$. Let $\hat{i} = \min\{\mathcal{I}(w)\}$, then

$$\begin{aligned} f_3(w) + \left\langle \mathsf{e}(\hat{i}), v - w \right\rangle &= w(\hat{i}) + \left\langle \mathsf{e}(\hat{i}), v - w \right\rangle \\ &= w(\hat{i}) + v(\hat{i}) - w(\hat{i}) \\ &= v(\hat{i}) \\ &\leq \max\left\{ \max_{i\in[d]}\{v(i)\}, 0 \right\}, \end{aligned}$$

as was to be shown.

Then, we provide analysis of the dynamics of GD using the first-order oracle described above. We only describe the dynamics under the event $\{T/2 \leq \|\mathsf{B}\|_0 \leq T\}$. Let the (ordered) set of bad coordinates denoted by $\mathcal{B} = \{i_1, \dots, i_{\|\mathsf{B}\|_0}\}$. The main observation here is that we can re-write the Eq. (16) as follows

$$\hat{\mathrm{F}}_{S_n}(w) = \sum_{i\in[d]\setminus\mathcal{B}} w(i)^2\hat{\mu}(i) + \lambda \sum_{i\in[d]\setminus\mathcal{B}} w(i)\hat{\mu}(i) + \max\left\{ \max_{i\in[d]}\{w(i)\}, 0 \right\}. \tag{18}$$

This equation shows that the gradient comes from the first two terms does not change the bad coordinates of w. As we will show that f_3 does not provide gradient for bad coordinates, the dynamic of each good coordinate of w is independent of other coordinates. Formally, we prove by induction that $W_1 = -\eta\lambda\hat{\mu}$, and for $t \geq 2$,

$$W_t(i) = \begin{cases} \frac{\lambda}{2}(-1 + (1 - 2\eta\hat{\mu}(i))^t) & i \in [d] \setminus \mathcal{B} \\ -\eta & i \in \{i_1, \ldots, i_{\min\{\|\mathsf{B}\|_0, t-1\}}\} \\ 0 & i \in \{i_{\min\{\|\mathsf{B}\|_0, t-1\}+1}, \ldots, i_{\|\mathsf{B}\|_0}\} \end{cases}.$$

For the base case, by the GD algorithm's update rule we have $W_1 = \Pi_{\mathcal{W}}\big(W_0 - \eta g_0\big) = \Pi_{\mathcal{W}}\big(-\eta g_0\big)$. Note that $g_0 = \lambda\hat{\mu}$. Since $\lambda \in \mathcal{O}\Big(1/(n\sqrt{d})\Big)$, $-\eta g_0 \in \mathcal{W}$.

For the inductive step, assume that for some $k \in [T-1]$, the claim holds. We have $W_{k+1} = \Pi_{\mathcal{W}}\big(W_k - \eta g_k\big)$. First, for $i \in [d] \setminus \mathcal{B}$, $g_k(i) = 2W_k(i)\hat{\mu}(i) + \lambda\hat{\mu}(i)$. Note that the gradient from the third term is zero for good coordinates as $W_k(i) < 0$ for $i \in [d] \setminus \mathcal{B}$. By a simple calculation, one can show that

$$W_k - \eta g_k = \frac{\lambda}{2}\big(-1 + (1 - 2\eta\hat{\mu}(i))^{k+1}\big).$$

Also, as $\lambda \in \mathcal{O}\Big(1/(n\sqrt{d})\Big)$, $W_k - \eta g_k \in \mathcal{W}$. Then, for the bad coordinates, consider two cases $\min\{\|\mathsf{B}\|_0\, k - 1\} = k - 1$ and $\min\{\|\mathsf{B}\|_0, k - 1\} = \|\mathsf{B}\|_0$. Consider the first case, i.e., $\min\{\|\mathsf{B}\|_0, k-1\} = k - 1$. Consider $i_k \in \mathcal{B}$. From Eq. (18), the first two terms do not provide gradient for bad coordinates. Then, we claim that $\partial f_3(W_k) = \mathsf{e}(i_k)$. The reason is that for all $i \in \{i_1, \ldots, i_{k-1}\} \cup [d] \setminus \mathcal{B}$, $W_k(i) < 0$ and $i \in \{i_k, \ldots, \|\mathsf{B}\|_0\}$, $W_k(i) = 0$. Therefore, the claim follows from Eq. (17). Therefore, for the first case, for all $i \in \{i_1, \ldots, i_k\}$, $W_{k+1}(i) = -\eta$, and for all $i \in \{i_{k+1}, \ldots, \|\mathsf{B}\|_0\}$, $W_{k+1}(i) = 0$.

Consider the second case, $\min\{\|\mathsf{B}\|_0\, k - 1\} = \|\mathsf{B}\|_0$. In this case, all coordinates of W_k are less than zero. Therefore, the gradient from f_3 is zero, and the bad coordinates remain unchanged. ■

Next, we provide a result regarding $\|W_T\|$.

Lemma 19 *Under the event* $\{T/2 \leq \|\mathsf{B}\|_0 \leq T\}$, *we have* $\frac{1}{2\sqrt{n}} \leq \|W_T\| \leq \frac{1}{\sqrt{n}}$.

Proof Under the event $\{T/2 \leq \|\mathsf{B}\|_0 \leq T\}$, Lemma 18 shows that

$$\|W_T\| = \Big(\|B\|_0\eta^2 + \sum_{i \in [d]\setminus\mathcal{B}} W_T(i)^2\Big)^{\frac{1}{2}}.$$

Since for good coordinates, $|W_T(i)| \leq \lambda\eta T$, we have the following upper bound $\|W_T\| \leq \sqrt{T\eta^2 + d(\lambda\eta T)^2}$. For a lower bound consider $\|W_T\| \geq \sqrt{T\eta^2/2}$. Setting the parameters, we obtain $\|W_T\| \leq \frac{1}{\sqrt{n}}$ and $\|W_T\| \geq \frac{1}{2\sqrt{n}}$. ■

C.3. Noise with Large Variance Fails

Consider the case that the variance of ξ along each dimension is σ^2 and $\sigma \geq \frac{\beta^\star}{\sqrt{d}}$ where $\beta^\star = 0.1$. In particular, $\frac{\beta^\star}{\sqrt{d}}$ is the threshold for the variance. First of all note that for all $w \in \mathcal{W}$

$$\mathrm{F}_{\mathcal{D}}(w) = \frac{1}{2}\|w\|^2 + \frac{\lambda}{2}\sum_{i=1}^{d} w(i) + \max\Big\{\max_{i \in [d]}\{w(i)\}, 0\Big\}.$$

Therefore,

$$|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| = \Big|\frac{1}{2}(\|\tilde{W}_T\|^2 - \|W\|^2) + \frac{\lambda}{2}\sum_{i=1}^{d}(\tilde{W}_T(i) - W_T(i)) + \Xi_T\Big|, \tag{19}$$

where $\Xi_T = \max\{\max_{i\in[d]}\{\tilde{W}_T(i)\}, 0\} - \max\{\max_{i\in[d]}\{W_T(i)\}, 0\}$. Under the event $\{T/2 \le \|\mathsf{B}\|_0 \le T\}$, Lemma 18 shows that $\max\{\max_{i\in[d]}\{W_T(i)\}, 0\} = 0$ since $W_T(i) < 0$ for all $i \in [d]$. Therefore, $\Xi_T = \max\{\max_{i\in[d]}\{\tilde{W}_T(i)\}, 0\} \ge 0$.

Because the Gaussian distribution is invariant under the rotation, we can assume that $W_T = (\|W_T\|, \underbrace{0, \dots, 0}_{d-1 \text{ times}})$ without loss of generality. Therefore, Eq. (19) is given by

$$|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| = \Big|\frac{1}{2}(\|\tilde{W}_T\|^2 - \|W_T\|^2) + \frac{\lambda}{2}(\tilde{W}_T(1) - \|W_T\| + \sum_{i=2}^{d}\tilde{W}_T(i)) + \Xi_T\Big|. \tag{20}$$

Let $V_T = W_T + \xi$. Let us represent $\xi = r\theta$ where $r = \|\xi\|$ and $\theta = \xi/\|\xi\|$. By a simple calculation, one can obtain that

$$\|V_T\|^2 = \|W_T\|^2 + r^2 + 2\|W_T\|r\theta(1). \tag{21}$$

Define $\mathsf{r}_{\max} = 1 - \|W_T\|$. Note $0 \le \mathsf{r}_{\max} \le 1$ since $W_T \in \mathcal{W}$. By the tower rule for the expectation,

$$\begin{aligned} \mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big] &\ge \mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\mathbb{1}[r \le \mathsf{r}_{\max}]\Big] \\ &+ \mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\mathbb{1}[r > \mathsf{r}_{\max}]\mathbb{1}[\|V_T\| \le 1]\Big] \\ &+ \mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\mathbb{1}[r > \mathsf{r}_{\max}]\mathbb{1}[\|V_T\| > 1]\Big] \end{aligned} \tag{22}$$

Under the event $\{T/2 \le \|\mathsf{B}\|_0 \le T\}$, we divide the sample space into three regions: **Region 1**: $\{r \le \mathsf{r}_{\max}\}$, **Region 2**: $\{r > \mathsf{r}_{\max}\} \cap \{\|V_T\| < 1\}$, and **Region 3**: $\{r > \mathsf{r}_{\max}\} \cap \{\|V_T\| \ge 1\}$. In what follows, we lower bound Eq. (22) for each region separately.

$[R_1]$ **Region 1** $\{r \le \mathsf{r}_{\max}\}$:

By the tower rule for the expectation,

$$\begin{aligned} &\mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\mathbb{1}[r \le \mathsf{r}_{\max}]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}^{W_T,r}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big]\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\mathbb{1}[r \le \mathsf{r}_{\max}]\Big] \end{aligned}$$

Under the event $\{r \le \mathsf{r}_{\max}\}$, it is straightforward to see that $\|V_T\| \le 1$. Therefore, $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T) = V_T$, and Eq. (20) is given by

$$|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| = \Big|\frac{1}{2}r^2 + \|W_T\|r\theta(1) + \frac{\lambda r}{2}\sum_{i=1}^{d}\theta(i) + \Xi_T\Big|.$$

By the construction of the surrogate algorithm and Lemma 21, we know that θ is independent of r and W_T. Then, we invoke the reverse triangle inequality, i.e., $|a-b| \geq |a| - |b|$ for $a, b \in \mathbb{R}$. Using $\theta \overset{d}{=} -\theta$, we have

$$\begin{aligned}
&\mathbb{E}^{r,W_T}\left[\left|\frac{1}{2}r^2 + r\|W_T\|\theta(1) + \frac{\lambda r}{2}\sum_{i=1}^{d}\theta(i) + \Xi_T\right|\right] \\
&\geq \underbrace{\mathbb{E}^{W_T,r}\left[\left|\frac{1}{2}r^2 + \|W_T\|r\theta(1) + \Xi_T\right|\right]}_{\textcircled{1}} - \underbrace{\mathbb{E}^{W_T,r}\left[\left|\frac{\lambda r}{2}\sum_{i=1}^{d}\theta(i)\right|\right]}_{\textcircled{2}}. \qquad (23)
\end{aligned}$$

We will analyze $\textcircled{1}$ and $\textcircled{2}$ separately. Note that $\theta(1) \sim \mathrm{Unif}([-1,1])$. Thus, with probability $1/2$, $\theta(1) \in [0,1]$. Therefore,

$$\textcircled{1} \geq \mathbb{E}^{W_T,r}\left[\left|\frac{1}{2}r^2 + \|W_T\|r\theta(1) + \Xi_T\right|\mathbb{1}[\theta(1) \in [0,1]]\right] \geq \left|\frac{r^2}{4} + \frac{\Xi_T}{2}\right| \geq \frac{r^2}{4}, \qquad (24)$$

where the last inequality follows from $\Xi_T \geq 0$. By the Cauchy-Schwartz, we have $\|\theta\|_1 \leq \sqrt{d}$ since $\|\theta\|_2 = 1$. Therefore,

$$\begin{aligned}
\textcircled{2} &\leq \frac{\lambda r}{2}\|\theta\|_1 \\
&\leq \frac{\lambda r}{2}\sqrt{d} \\
&\leq \frac{\lambda \mathsf{r}_{\max}}{2}\sqrt{d} \\
&\leq \frac{\lambda}{2}\sqrt{d}, \qquad (25)
\end{aligned}$$

where the third inequality follows since $r \leq \mathsf{r}_{\max}$ and the the last step follows from $\mathsf{r}_{\max}$ being less than one. By Eq. (23), Eq. (24), and Eq. (25), we finish lower bounding the inner expectation,

$$\mathbb{E}^{W_T,r}\left[|\mathsf{F}_{\mathcal{D}}(\tilde{W}_T) - \mathsf{F}_{\mathcal{D}}(W_T)|\right] \geq \frac{r^2}{4} - \frac{\lambda\sqrt{d}}{2} \geq \frac{r^2}{4} - \frac{1}{2n}. \qquad (26)$$

Here, the last inequality follows from setting $\lambda \leq \frac{1}{n\sqrt{d}}$.

$[R_2]$ **Region 2**: $\{r > \mathsf{r}_{\max}\}$ and $\{\|V_T\| < 1\}$:

Since $\|V_T\| < 1$ and $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T)$, we have $\tilde{W}_T = V_T$. Using Eq. (21), we can write

$$|\mathsf{F}_{\mathcal{D}}(\tilde{W}_T) - \mathsf{F}_{\mathcal{D}}(W_T)| \geq |\frac{1}{2}(r^2 + 2\|W_T\|r\theta(1)) + \frac{\lambda r}{2}\sum_{i\in[d]}\theta(i) + \Xi_T|.$$

Then, using the reverse triangle inequality, i.e., $|a-b| \geq |a| - |b|$ for $a, b \in \mathbb{R}$, and the facts that $|\theta(1)| \leq 1$ and $\Xi_T \geq 0$, we have

$$\begin{aligned}|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| &\geq \Big|\frac{1}{2}r^2 + \Xi_T\Big| - \Big|r\|W_T\|\theta(1) + \frac{\lambda r}{2}\sum_{i\in[d]}\theta(i)\Big| \\ &\geq \Big|\frac{1}{2}r^2 + \Xi_T\Big| - r\|W_T\||\theta(1)| - \frac{\lambda r}{2}\Big|\sum_{i\in[d]}\theta(i)\Big| \\ &\geq \frac{1}{2}r^2 + \Xi_T - r\|W_T\| - \frac{\lambda r}{2}\Big|\sum_{i\in[d]}\theta(i)\Big| \\ &\geq \frac{1}{2}r^2 - r\|W_T\| - \frac{\lambda r}{2}\Big|\sum_{i\in[d]}\theta(i)\Big|.\end{aligned}$$

By the Cauchy-Schwartz, we have $\|\theta\|_1 \leq \sqrt{d}$ since $\|\theta\|_2 = 1$. Therefore,

$$\begin{aligned}|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| &\geq \frac{1}{2}r^2 - r\|W_T\| - \frac{\lambda r}{2}\sqrt{d} \\ &\geq \frac{1}{2}r^2 - r\|W_T\| - \frac{r}{2n}. \qquad (27)\end{aligned}$$

Here, the last line follows from $\lambda \leq \frac{1}{n\sqrt{d}}$.

Define $g : \mathbb{R} \to \mathbb{R}$ where $g(x) = x^2/2 - x(\|W_T\| + 1/(2n))$. Then, we have $\arg\min_{x\in\mathbb{R}} g(x) = \|W_T\| + 1/(2n)$. From Lemma 19, we know that $\|W_T\| \leq 1/\sqrt{n}$. Notice that for $n \geq 5$, we have $\|W_T\| \leq 1/\sqrt{n} \leq 0.5(1-1/(2n))$ which gives us $\|W_T\| + 1/(2n) \leq 1 - \|W_T\| = \mathsf{r}_{\max}$. Therefore, we conclude that g is increasing for $x \geq \mathsf{r}_{\max}$. Note that the lower bound in Eq. (27) is $g(r)$ and using this observation we have $g(r) > g(\mathsf{r}_{\max})$ since in this region $r > \mathsf{r}_{\max}$. Therefore, we can further lower bound Eq. (27) as

$$\begin{aligned}|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| &\geq \frac{3}{2}\|W_T\|^2 - \Big(2 - \frac{1}{2n}\Big)\|W_T\| + \frac{1}{2}\Big(1 - \frac{1}{n}\Big) \\ &\geq \frac{1}{2} - \frac{2}{\sqrt{n}}. \qquad (28)\end{aligned}$$

To prove the last step define $h : \mathbb{R} \to \mathbb{R}$ where $h(x) = \frac{3}{2}x^2 - \Big(2 - \frac{1}{2n}\Big)x + \frac{1}{2}\Big(1 - \frac{1}{n}\Big)$. It is straightforward to see that $h(x)$ is decreasing for $x \leq 1/\sqrt{n}$ when $n \geq \sqrt{5}$. Using this argument and some manipulations we can show the last step.

$[R_3]$ **Region 3**: $\{r > \mathsf{r}_{\max}\}$ and $\{\|V_T\| \geq 1\}$.

Since $\|V_T\| \geq 1$ and $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T)$, we have $\|\tilde{W}_T\| = 1$. Using this observation and reverse triangle inequality, i.e., $|a-b| \geq |a| - |b|$ for $a, b \in \mathbb{R}$, we can simplify Eq. (20) as

$$\begin{aligned}|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| &\geq \frac{1}{2}|1 - \|W_T\|^2 + 2\Xi_T| - \frac{\lambda}{2}\Big|\sum_{i\in[d]}\tilde{W}_T(i) - \|W_T\|\Big| \\ &\geq \frac{1}{2}|1 - \|W_T\|^2 + 2\Xi_T| - \frac{\lambda}{2}(\|\tilde{W}_T\|_1 + \|W_T\|).\end{aligned}$$

The last line follows from using the triangle inequality twice. By Lemma 19, we have $\|W_T\| \leq 1/\sqrt{n}$. Then, since $\Xi_T \geq 0$, we obtain

$$|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| \geq \frac{1}{2}(1 + 2\Xi_T - \frac{1}{n}) - \frac{\lambda}{2}(\|\tilde{W}_T\|_1 + \frac{1}{\sqrt{n}}).$$

Also, by the Cauchy-Schwartz, we have $\|\tilde{W}_T\|_1 \leq \sqrt{d}$ since $\|\tilde{W}_T\|_2 = 1$. Therefore, setting $\lambda \leq \frac{1}{n\sqrt{d}}$

$$\begin{aligned}|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)| &\geq \frac{1}{2}(1 - \frac{1}{n}) + \Xi_T - \frac{1}{2n} - \frac{1}{2n^{1.5}\sqrt{d}} \\ &\geq \frac{1}{2} - \frac{1}{n} \\ &\geq \frac{1}{2} - \frac{2}{\sqrt{n}}. \qquad (29)\end{aligned}$$

Here the last line follows from $\Xi_T \geq 0$ and some simple manipulations.

Equipped with the lower bounds for each region we can conclude this part of the proof. Combining Eq. (22) with Eq. (26), Eq. (28), and Eq. (29), we obtain

$$\begin{aligned}\mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big] &\geq \mathbb{E}\Big[\Big(\frac{r^2}{4} - \frac{1}{2n}\Big)\mathbb{1}[T/2 \leq \|\mathsf{B}\|_0 \leq T]\mathbb{1}[r \leq \mathsf{r}_{\max}]\Big] \\ &+ \Big(\frac{1}{2} - \frac{2}{\sqrt{n}}\Big)\mathbb{E}\big[\mathbb{P}^{S_n}(r > \mathsf{r}_{\max})\mathbb{1}[T/2 \leq \|\mathsf{B}\|_0 \leq T]\big]. \qquad (30)\end{aligned}$$

Assume we choose n sufficiently large so that $\frac{(\beta^\star)^2}{16} - \frac{1}{2n} \geq 0$ (Notice that such n always exists). We can further lower bound Eq. (30) as

$$\begin{aligned}&\mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big] \geq \mathbb{E}\Big[\Big(\frac{r^2}{4} - \frac{1}{2n}\Big)\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\mathbb{1}[\frac{\beta^\star}{2} \leq r \leq \mathsf{r}_{\max}]\Big] \\ &+ \mathbb{E}\Big[\Big(\frac{r^2}{4} - \frac{1}{2n}\Big)\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\mathbb{1}[r < \frac{\beta^\star}{2}]\Big] \\ &+ \Big(\frac{1}{2} - \frac{2}{\sqrt{n}}\Big)\mathbb{E}\Big[\mathbb{P}^{S_n}(r > \mathsf{r}_{\max})\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\Big] \\ &\geq \Big(\frac{(\beta^\star)^2}{16} - \frac{1}{2n}\Big)\mathbb{E}\Big[\mathbb{P}^{S_n}(\frac{\beta^\star}{2} \leq r \leq \mathsf{r}_{\max})\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\Big] \\ &- \frac{1}{2n}\mathbb{E}\Big[\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\mathbb{1}[r < \frac{\beta^\star}{2}]\Big] + \Big(\frac{1}{2} - \frac{2}{\sqrt{n}}\Big)\mathbb{E}\Big[\mathbb{P}^{S_n}(r > \mathsf{r}_{\max})\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\Big] \\ &\geq \Big(\frac{(\beta^\star)^2}{16} - \frac{1}{2n}\Big)\mathbb{E}\Big[\mathbb{P}^{S_n}(\frac{\beta^\star}{2} \leq r)\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\Big] - \frac{1}{2n}\mathbb{E}\Big[\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\mathbb{1}[r < \frac{\beta^\star}{2}]\Big] \\ &\geq \Big(\frac{(\beta^\star)^2}{16} - \frac{1}{2n}\Big)\mathbb{E}\Big[\mathbb{P}^{S_n}(\frac{\beta^\star}{2} \leq r)\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\Big] - \frac{1}{2n}\mathbb{P}(r < \frac{\beta^\star}{2}). \qquad (31)\end{aligned}$$

Here, we have used $\frac{1}{2} - \frac{2}{\sqrt{n}} \geq \frac{(\beta^\star)^2}{16} - \frac{1}{2n}$ for $n \geq 14$ where $\beta^\star = 0.1$, and $\frac{r^2}{4} - \frac{1}{2n} \geq -\frac{1}{2n}$ for $r \geq 0$.

Note that $S_n \perp\!\!\!\perp r$ by the construction of the surrogate algorithm. By assumption we have $\sigma \geq \frac{\beta^\star}{\sqrt{d}}$. Using the concentration bound from Corollary 23 we obtain

$$\mathbb{P}\Big(r \leq \frac{\beta^\star}{2}\Big) \leq \mathbb{P}\Big(r \leq \frac{\sigma\sqrt{d}}{2}\Big) \leq 2\exp\Big(-\frac{9d}{64}\Big).$$

Since r and S_n are independent, by Eq. (15) we have

$$\begin{aligned}\mathbb{E}\Big[\mathbb{P}^{S_n}(\frac{\beta^\star}{2} \leq r)\mathbb{1}[\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T]\Big] &= \mathbb{P}\Big(\frac{\beta^\star}{2} \leq r\Big)\mathbb{P}\Big(\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T\Big) \\ &\geq (1 - 2\exp(-9d/64))\mathbb{P}\Big(\frac{T}{2} \leq \|\mathsf{B}\|_0 \leq T\Big) \\ &\geq (1 - 2\exp(-9d/64))(1 - 2\exp(-T/36)). \qquad (32)\end{aligned}$$

Therefore, we conclude this part by combining Eq. (31) and Eq. (32) to obtain the following lower bound:

$$\begin{aligned}&\mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big] \\ &\geq \Big(\frac{(\beta^\star)^2}{16} - \frac{1}{2n}\Big)\big(1 - 2\exp(-9d/16) - 2\exp(-T/36)\big) - \frac{1}{n}\exp(-\frac{9d}{64}).\end{aligned}$$

By setting the parameters, i.e., T and d, we prove that for sufficiently large n

$$\mathbb{E}\Big[|\mathrm{F}_{\mathcal{D}}(\tilde{W}_T) - \mathrm{F}_{\mathcal{D}}(W_T)|\Big] \in \Omega(1),$$

which was to be shown.

C.4. Noise With Small Variance Fails: IOMI

In Appendix C.3 we showed that if the variance of ξ is greater than $\big(\frac{\beta^\star}{\sqrt{d}}\big)^2$, the distance between the population risk of the surrogate algorithm and the GD algorithm does not go to zero. In this part, we will show that if the variance of ξ is smaller than $\frac{\beta^\star}{\sqrt{d}}$ then, the mutual information term does not vanish as $n \to \infty$.

By the definition of the mutual information we can write $I(\tilde{W}_T; S_n) = \mathrm{H}(S_n) - \mathrm{H}(S_n|\tilde{W}_T)$. Note that B is a S_n-measurable random variable. Therefore, we have

$$\mathrm{H}(S_n|\tilde{W}_T) = \mathrm{H}(S_n, \mathsf{B}|\tilde{W}_T).$$

Then, by the chain rule for the discrete entropy $\mathrm{H}(S_n, \mathsf{B}|\tilde{W}_T) = \mathrm{H}(\mathsf{B}|\tilde{W}_T) + \mathrm{H}(S_n|\tilde{W}_T, \mathsf{B})$. We claim that

$$\mathrm{H}(S_n|\tilde{W}_T, \mathsf{B}) \leq n\mathbb{E}[(d - \|\mathsf{B}\|_0)].$$

The reason is by conditioning on B, we know the exact values for the bad coordinates in S_n. Therefore, the cardinally of the possible values for each data-point, conditioned on B, cannot be more that $2^{d-\|\mathsf{B}\|_0}$. Thus,

$$\begin{aligned}I(\tilde{W}_T; S_n) &\geq \mathrm{H}(S_n) - \mathrm{H}(\mathsf{B}|\tilde{W}_T) - n(d - \mathbb{E}[\|\mathsf{B}\|_0]) \\ &= n\mathbb{E}[\|\mathsf{B}\|_0] - \mathrm{H}(\mathsf{B}|\tilde{W}_T),\end{aligned}$$

where the last line follows from $\mathrm{H}(S_n) = nd$ because each element in S_n is drawn i.i.d. Also, note that $\mathbb{E}[\|\mathsf{B}\|_0] = \mathbb{E}[\sum_{i=1}^{d} \mathsf{B}(i)] = d\mathbb{E}[\mathsf{B}(1)] = d2^{-n}$ where in the last line we used the fact that each element of S_n is i.i.d., and each column is a bad coordinate with probability 2^{-n}. Also, with the similar reasoning we obtain $\mathrm{H}(\mathsf{B}) = \mathrm{H}(\mathsf{B}(1), \ldots, \mathsf{B}(d)) = d\mathrm{H}_b(2^{-n})$, where for $x \in [0,1]$ $\mathrm{H}_b(x) = -x\log(x) - (1-x)\log(1-x)$ is the binary entropy function.

Then, we invoke a version of Fano's inequality, provided in Lemma 24, to obtain

$$\mathrm{H}(\mathsf{B}|\tilde{W}_T) \le 1 + \mathsf{P_e}\mathrm{H}(\mathsf{B})$$

where $\mathsf{P_e} = \inf_{M:\mathcal{W}\to\{0,1\}^d} \mathbb{P}(M(\tilde{W}_T) \neq \mathsf{B})$. Using the well-known inequality $\mathrm{H}_b(x) \le -x\log(x) + x$, we obtain

$$\mathrm{H}(\mathsf{B}) \le d(n2^{-n} + 2^{-n}) = d(n+1)2^{-n}.$$

Therefore,

$$\begin{aligned} I(\tilde{W}_T; S) &\ge nd2^{-n} - (n+1)d2^{-n}\mathsf{P_e} - 1 \\ &= nd2^{-n}\Big(1 - \frac{n+1}{n}\mathsf{P_e}\Big) - 1 \\ &\ge 1.5n^3(1 - 2\mathsf{P_e}) - 1, \end{aligned} \tag{33}$$

where the last line follows from setting $d = 0.75T2^n$, $T = 2n^2$, and $(n+1)/n \le 2$.

Next, we design an estimator Ψ to *decode* B from $\tilde{W}_T$ and analyze its probability of error. Let $h = (\eta + \eta\lambda T)/2$. Then, the proposed estimator is given by

$$\Psi(w)(i) = \begin{cases} 1 & \text{if } |w(i)| \ge h \\ 0 & \text{if } |w(i)| < h \end{cases} \tag{34}$$

for $i \in [d]$. In words, it compares each coordinate of w with a given threshold, and if it is larger than h, then that coordinate declares as a bad coordinate.

Let $V_T = W_T + \xi$. Then,

$$\begin{aligned} \mathsf{P_e} &\le \mathbb{P}(\exists i \in [d] \text{ s.t. } \Psi(\tilde{W}_T)(i) \neq \mathsf{B}(i)) \\ &\le \mathbb{P}(\{\exists i \in [d] \text{ s.t. } \Psi(\tilde{W}_T)(i) \neq \mathsf{B}(i)\} \cap \{\|V_T\| \le 1\}) + \mathbb{P}(\|V_T\| \ge 1) \end{aligned} \tag{35}$$

First we show that $\mathbb{P}(\|V_T\| \ge 1)$ is sufficiently small. From Eq. (21), we have $\|V_T\| \ge 1 = \|W_T\|^2 + r^2 + 2\|W_T\|r\theta(1)$. Then, as shown in Appendix C.3 given that $\{r \le \mathsf{r}_{\mathsf{max}} = 1 - \|W_T\|\}$, then $\|V_T\| \le 1$. Using this we obtain

$$\begin{aligned} \mathbb{P}(\|V_T\| \ge 1) &= \mathbb{P}(\|W_T\|^2 + r^2 + 2\|W_T\|r\theta(1) \ge 1) \\ &\le \mathbb{P}(r \ge 1 - \|W_T\|). \end{aligned}$$

Here $\xi = r\theta$ where $r = \|\xi\|$ and $\theta = \xi/\|\xi\|$. Recall from Lemma 19 that under the event $\{T/2 \le \|\mathsf{B}\|_0 \le T\}$, $1/(2\sqrt{n}) \le \|W_T\| \le 1/\sqrt{n}$. Therefore,

$$\begin{aligned} \mathbb{P}(r \ge 1 - \|W_T\|) &\le \mathbb{E}\big[\mathbb{P}^{S_n}(r \ge 1 - \|W_T\|)\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\big] + 1 - \mathbb{P}(T/2 \le \|\mathsf{B}\|_0 \le T) \\ &\le \mathbb{E}\big[\mathbb{P}^{S_n}(r \ge 1 - 1/(2\sqrt{n}))\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\big] + 1 - \mathbb{P}(T/2 \le \|\mathsf{B}\|_0 \le T) \\ &\le \mathbb{E}\big[\mathbb{P}^{S_n}(r \ge 1 - 1/(2\sqrt{n}))\mathbb{1}[T/2 \le \|\mathsf{B}\|_0 \le T]\big] + 2\exp(-T/36), \end{aligned}$$

where the last line follows from Eq. (15). Observe that

$$\{r \geq 1 - 1/(2\sqrt{n})\} \subseteq \{r \geq 2\beta^\star\},$$

due to $\beta^\star = 0.1$. Also as $\sigma \leq \beta^\star/\sqrt{d}$, we have

$$\mathbb{P}(r \geq 2\beta^\star) \leq \mathbb{P}(r \geq \sqrt{4d(\sigma^\star)^2}) \leq 2\exp\Big(-\frac{9d}{16}\Big),$$

where the last inequality comes from the concentration bounds for r in Corollary 23. Since $r \perp\!\!\!\perp S_n$, we have

$$\mathbb{P}^{S_n}(r \geq 1 - 1/(2\sqrt{n})) \leq 2\exp\Big(-\frac{9d}{16}\Big).$$

Therefore,

$$\mathbb{P}(\|V_T\| \geq 1) \leq 2\exp\Big(-\frac{9d}{16}\Big) + 2\exp(-T/36). \tag{36}$$

Since under the event $\|V_T\| \leq 1$, $\tilde{W}_T = \Pi_{\mathcal{W}}(V_T) = W_T + \xi$,

$$\begin{aligned}
&\mathbb{P}(\{\exists i \in [d] \text{ s.t. } \Psi(\tilde{W}_T)(i) \neq \mathsf{B}(i)\} \cap \{\|V_T\| \leq 1\}) \\
&= \mathbb{P}(\{\exists i \in [d] \text{ s.t. } \Psi(W_T + \xi)(i) \neq \mathsf{B}(i)\} \cap \{\|V_T\| \leq 1\}) \\
&\leq \mathbb{P}(\exists i \in [d] \text{ s.t. } \Psi(W_T + \xi)(i) \neq \mathsf{B}(i)\}).
\end{aligned} \tag{37}$$

By the definition of the error probability

$$\mathbb{P}(\forall i \in [d]\ \Psi(W_T + \xi)(i) = \mathsf{B}(i)) \geq \mathbb{E}[\mathbb{P}^{S_n}(\forall i \in [d]\ \Psi(W_T + \xi)(i) = \mathsf{B}(i))\mathbb{1}[T/2 \leq \|\mathsf{B}\|_0 \leq T]]. \tag{38}$$

Note that W_T and B are S_n-measurable. Therefore, the inner probability is only over ξ. Also, let $\mathcal{B} = \{i_1, \dots, i_{\|\mathsf{B}\|_0}\}$ denote the set of bad coordinates. Using the closed-form expression in Lemma 18 for W_T under the event $\{T/2 \leq \|\mathsf{B}\| \leq T\}$, we have

$$\mathbb{P}^{S_n}(\forall i \in [d]\ \Psi(W_T + \xi)(i) = \mathsf{B}(i)) = \big(\Pi_{i\in\mathcal{B}}\mathbb{P}^{S_n}(\eta + \xi(i) \geq h)\big)\,\Pi_{i\in[d]\setminus\mathcal{B}}\left(\mathbb{P}^{S_n}(-W_T(i) + \xi(i) \leq h)\right). \tag{39}$$

This identity follows from $\xi \perp\!\!\!\perp S_n$ and each coordinate of ξ are i.i.d. As shown in Lemma 18, under the event $\{T/2 \leq \|\mathsf{B}\| \leq T\}$, $0 \leq -W_T(i) \leq \lambda\eta T$; therefore, $-W_T(i) < h$ for $i \in [d] \setminus \mathsf{B}$. We can simplify Eq. (39) as

$$\begin{aligned}
&\big(\Pi_{i\in\mathcal{B}}\mathbb{P}^{S_n}(\eta + \xi(i) \geq h)\big)\,\Pi_{i\in[d]\setminus\mathcal{B}}\left(\mathbb{P}^{S_n}(-W_T(i) + \xi(i) \leq h)\right) \\
&= \Big(1 - Q\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big)\Big)^{\|\mathsf{B}\|_0}\Pi_{i\in[d]\setminus\mathcal{B}}\Big(1 - Q\Big(\frac{h + W_T(i)}{\sigma^\star}\Big)\Big)
\end{aligned}$$

where for $x \in \mathbb{R}$, $Q(x) = \frac{1}{\sqrt{2\pi}}\int_{t\geq x}\exp(-\frac{t^2}{2})\mathrm{d}t$ is the tail distribution function of the Gaussian distribution with mean zero and variance one. Since $Q\Big(\frac{h+W_T(i)}{\sigma^\star}\Big) \leq Q\Big(\frac{h-\eta\lambda T}{\sigma^\star}\Big) = Q\Big(\frac{\eta-\eta\lambda T}{2\sigma^\star}\Big)$ for all $i \in [d] \setminus \mathcal{B}$, we can further lower bound as

$$\begin{aligned}
\mathbb{P}^{S_n}(\forall i \in [d]\ \Psi(W_T + \xi)(i) = \mathsf{B}(i)) &\geq \big(\Pi_{i\in\mathcal{B}}\mathbb{P}^{S_n}(\eta + \xi(i) \geq h)\big)\,\Pi_{i\in[d]\setminus\mathcal{B}}\left(\mathbb{P}^{S_n}(\eta\lambda T + \xi(i) \leq h)\right) \\
&\geq \Big(1 - Q\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big)\Big)^{\|\mathsf{B}\|_0}\Big(1 - Q\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big)\Big)^{d-\|\mathsf{B}\|_0},
\end{aligned} \tag{40}$$

More precisely, since $\eta\lambda T < h < \eta$, we have $\mathbb{P}^{S_n}(\eta + \xi(i) \geq h) = \mathbb{P}^{S_n}(\xi(i) \geq h - \eta) = 1 - \mathbb{P}^{S_n}(\xi(i) \geq \eta - h)$ and $\mathbb{P}^{S_n}(\eta\lambda T + \xi(i) \leq h) = \mathbb{P}^{S_n}(\xi(i) \leq h - \eta\lambda T) = 1 - \mathbb{P}^{S_n}(\xi(i) \geq h - \eta\lambda T)$.

Therefore, we can use Eq. (38), Eq. (39), and Eq. (40) to obtain

$$\begin{aligned}\mathbb{P}(\forall i \in [d]\ \Psi(W_T + \xi)(i) = \mathsf{B}(i)) &\geq \Big(1 - Q\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big)\Big)^d\ \mathbb{P}(T/2 \leq \|\mathsf{B}\|_0 \leq T)\\ &\geq \Big(1 - Q\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big)\Big)^d\ \Big(1 - 2\exp\Big(-\frac{T}{36}\Big)\Big),\end{aligned}$$

where in the last line we have used Eq. (15). Note that $\eta - \eta\lambda T \geq 0$ since $\lambda \in \mathcal{O}\Big(1/(n\sqrt{d})\Big)$. We can use the well-known inequality $(1-x)^n \geq 1 - nx$ for $x \leq 1$, $n \in \mathbb{N}$ to obtain

$$\begin{aligned}&1 - \mathbb{P}(\forall i \in [d]\ \Psi(W_T + \xi)(i) = \mathsf{B}(i))\\ &\leq dQ\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big) + 2\exp\Big(-\frac{T}{36}\Big) - 2dQ\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big)\exp\Big(-\frac{T}{36}\Big)\\ &\leq dQ\Big(\frac{\eta - \eta\lambda T}{2\sigma^\star}\Big) + 2\exp\Big(-\frac{T}{36}\Big),\end{aligned}$$

Then, we invoke the inequality $Q(x) \leq \frac{1}{2}\exp(-\frac{x^2}{2})$ for $x \geq 0$ [**wainwright2019high**], to further upper bound the last equation as follows:

$$1 - \mathbb{P}(\forall i \in [d]\ \Psi(W_T + \xi)(i) = \mathsf{B}(i)) \leq \frac{d}{2}\exp\Big(-\frac{d(\eta - \eta\lambda T)^2}{2(\beta^\star)^2}\Big) + 2\exp\Big(-\frac{T}{36}\Big). \qquad (41)$$

Finally, by combining Eq. (35), Eq. (36), Eq. (37), and Eq. (41), we obtain

$$\mathsf{P_e} \leq \frac{d}{2}\exp\Big(-\frac{d(\eta - \eta\lambda T)^2}{2(\beta^\star)^2}\Big) + 4\exp\Big(-\frac{T}{36}\Big) + 2\exp\Big(-\frac{9d}{16}\Big).$$

By setting the parameters and some simple manipulations, we obtain

$$\mathsf{P_e} \leq n^2 2^n \exp(-2^n/n) + 6\exp(-n^2/18). \qquad (42)$$

In Fig. 1, we plot the upper bound in Eq. (42). As can be seen the upper bound is decreasing and smaller than 0.1 for $n \geq 10$.

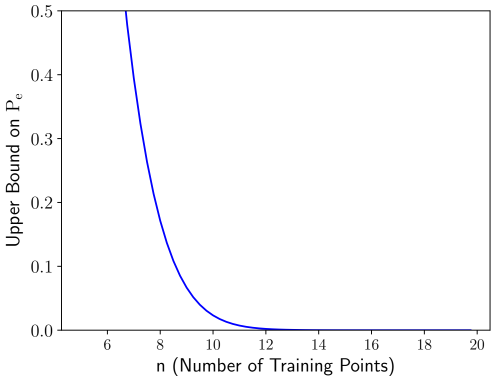

Figure 1: The upper bound in Eq. (42)

Finally, combining Eq. (42) with Eq. (33), we conclude that for $n \geq 10$ if $\sigma \leq \beta^\star/\sqrt{d}$, we have

$$I(\tilde{W}_T; S_n) \geq 1.2n^3 - 1,$$

which was to be shown.

C.5. Noise with Small Variance Fails: CMI

In this part of the proof we aim to show that if the variance of the noise is smaller than $\frac{(\beta^\star)^2}{d}$, then $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ grows linearly with n. We begin this part of the proof with a key lemma.

We recall the definition of bad coordinates. For $i \in [d]$, we say the $i-$th coordinate is a *bad coordinate* iff for all $j \in [n]$, $Z_j(i) = 0$. In words, if $i-$th coordinate is a bad coordinate then all the entries in the $i-$th column of S_n is zero. Let $\mathsf{B} \in \{0,1\}^d$ denote a vector such that $\mathsf{B}(i) = 1$ if and only if i is a bad coordinate. Also $\|\mathsf{B}\|_0$ denotes the number of bad coordinates.

Next, we provide a result which shows that U can be identified with high accuracy by having access to the supersample and bad coordinates. The intuition behind the result is as follows. Consider a decision making problem where by having access to B and matrix of $\tilde{S}$, we want to find which subset of the supersample contained in the training set, i.e., find U. First, note that, by definition, in each column of $\tilde{S}$ exactly one sample is chosen for the training set. Also, by the definition of the bad coordinates, we know that if $i \in [d]$ is a bad coordinate, then for all $Z \in S_n$, we have $Z(i) = 0$. In the next theorem we show that the *uncertainty* about U is small conditioned on $\mathsf{B}, \tilde{S}$. The idea of the proof is to show that by only considering the bad coordinates we can *distinguish* between the points in each column of the supersample.

Lemma 20 $\mathrm{H}(U|\mathsf{B}, \tilde{S}) \leq n\mathbb{E}[2^{-\|\mathsf{B}\|_0}]$.

Proof Let $\mathcal{B} = \{i_1, \ldots, i_{\|\mathsf{B}\|_0}\} \subseteq [d]$ contains the *ordered* set of bad coordinates. For every $k \in [n]$, define the following indicator random variable

$$J_k = \mathbb{1}[\exists i \in \mathcal{B} \text{ s.t. } \tilde{Z}_{0,k}(i) \neq \tilde{Z}_{1,k}(i)]$$

Let $J = (J_1, \ldots, J_n) \in \{0,1\}^n$. Note that J is $(\tilde{S}, \mathsf{B})$-measurable.

The main observation here is that provided that $J_k = 1$, then we can perfectly recover U_k. The reason is as follows: in each column of $\tilde{S}$, exactly one sample is a member of the training set. Also, since we know B, the values of the bad coordinates are known for the points in the training set by the definition of bad coordinates. Therefore, $J_k = 1$ iff one of the point in the k-th column of $\tilde{S}$ does not have zero on the indices in $\mathcal{B}$, which reveals the sample that is not in the training set. Therefore, as J is $(\tilde{S}, \mathsf{B})$-measurable, we can write

$$\begin{aligned}
\mathrm{H}(U|\mathsf{B}, \tilde{S}) &= \mathrm{H}(U|\mathsf{B}, \tilde{S}, J) \\
&= \mathrm{H}((U)_{\{i|J_i=0\}}, (U)_{\{i|J_i=1\}}|\mathsf{B}, \tilde{S}, J) \\
&= \mathrm{H}((U)_{\{i|J_i=0\}}|\mathsf{B}, \tilde{S}, J),
\end{aligned}$$

where the last line follows from $(U)_{\{i|J_i=1\}}$ being known from J. Since the cardinality of the support of $(U)_{\{i|J_i=0\}}$ is no more than $2^{n-\|J\|_0}$, we obtain

$$\mathrm{H}(U|\mathsf{B}, \tilde{S}) \leq n - \mathbb{E}[\|J\|_0].$$

Then, we claim that

$$\mathbb{P}(J_k = 1) = \mathbb{E}[1 - 2^{-\|\mathsf{B}\|_0}].$$

This claim conclude the proof since $\mathbb{E}[\|J\|_0] = \sum_{k=1}^{n} \mathbb{E}[J_k] = \sum_{k=1}^{n} \mathbb{P}(J_k = 1)$.

To prove the claim: $J_k = 0$ iff, conditioned on the U and B, for all j such that $\mathsf{B}_j = 1$, $Z_{1-U_k,k}(j) = 0$. By the definition of the supersample, the points in the supersample are i.i.d. , independent of U, and drawn from $\mathrm{Ber}(1/2)$. Hence,

$$\mathbb{P}(J_k = 0) = \mathbb{E}[\mathbb{P}^{U,S_n}(J_k = 0)] = \mathbb{E}[2^{-\|\mathsf{B}\|_0}].$$

■

By the definition of mutual information, we have

$$\begin{aligned}\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= \mathrm{H}(U|\tilde{S}) - \mathrm{H}(U|\tilde{W}_T, \tilde{S}) \\ &= \mathrm{H}(U) - \mathrm{H}(U|\tilde{W}_T, \tilde{S}) \\ &= n - \mathrm{H}(U|\tilde{W}_T, \tilde{S}), \end{aligned} \tag{43}$$

where the second and third steps follow from $U \perp\!\!\!\perp \tilde{S}$ and $\mathrm{H}(U) = n$, respectively. To analyze the second term in Eq. (43), consider the following equality which comes from the chain rule:

$$\begin{aligned}\mathrm{H}(U, \mathsf{B}|\tilde{W}_T, \tilde{S}) &= \mathrm{H}(U|\tilde{W}_T, \tilde{S}) + \mathrm{H}(\mathsf{B}|U, \tilde{W}_T, \tilde{S}) \\ &= \mathrm{H}(\mathsf{B}|\tilde{W}_T, \tilde{S}) + \mathrm{H}(U|\tilde{W}_T, \tilde{S}, \mathsf{B}).\end{aligned}$$

Notice that $\mathrm{H}(B|U, \tilde{W}_T, \tilde{S}) = 0$ as B is $(U, \tilde{S})$-measurable. Therefore,

$$\mathrm{H}(U|\tilde{W}_T, \tilde{S}) = \mathrm{H}(\mathsf{B}|\tilde{W}_T, \tilde{S}) + \mathrm{H}(U|\tilde{W}_T, \tilde{S}, \mathsf{B}). \tag{44}$$

To analyze the first term, note that conditioning cannot increase the entropy. Therefore, we have $\mathrm{H}(\mathsf{B}|\tilde{W}_T, \tilde{S}) \leq \mathrm{H}(\mathsf{B}|\tilde{W}_T)$. Then, we invoke the Fano's inequality from Lemma 24 to obtain

$$\mathrm{H}(\mathsf{B}|\tilde{W}_T) \leq 1 + \mathsf{P_e}\mathrm{H}(\mathsf{B}).$$

Here, $\mathsf{P_e} = \inf_{M:\mathcal{W}\to\{0,1\}^d} \mathbb{P}(M(\tilde{W}_T) \neq \mathsf{B})$. Consider the estimator Ψ proposed in Eq. (34). We analyzed its probability of error in Appendix C.4 and obtained in Eq. (42) that

$$\mathsf{P_e} \leq n^2 2^n \exp(-2^n/n) + 6\exp(-n^2/18).$$

Note that $\mathrm{H}(\mathsf{B}) \leq d(n+1)2^{-n} \leq 2n^3$ for $n \geq 3$ as shown in Appendix C.4. Therefore,

$$\mathrm{H}(\mathsf{B}|\tilde{W}_T, \tilde{S}) \leq \mathrm{H}(\mathsf{B}|\tilde{W}_T) \leq 2n^3(n^2 2^n \exp(-2^n/n) + 6\exp(-n^2/18)) + 1. \tag{45}$$

Next, we analyze the second term in Eq. (44). Using Lemma 20 we have

$$\mathrm{H}(U|\tilde{W}_T, \tilde{S}, \mathsf{B}) \leq \mathrm{H}(U|\tilde{S}, \mathsf{B}) \leq n\mathbb{E}[2^{-\|\mathsf{B}\|_0}]. \tag{46}$$

The, consider

$$\mathbb{E}[2^{-\|\mathsf{B}\|_0}] = \mathbb{E}[2^{-\|\mathsf{B}\|_0}\mathbb{1}[T/2 \leq \|\mathsf{B}\|_0 \leq T]] + \mathbb{E}[2^{-\|\mathsf{B}\|_0}(\mathbb{1}[\|\mathsf{B}\|_0 < T/2] + \mathbb{1}[\|\mathsf{B}\|_0 > T])].$$

The second term can be upper bounded by $\mathbb{P}(\{\|\mathsf{B}\|_0 < T/2\} \cup \{\|\mathsf{B}\|_0 > T\})$, and this probability is less than $2\exp(-T/36)$ as shown in Eq. (15). By simply upper bounding the first term by the worst-case realization, we can write

$$\begin{aligned}\mathbb{E}[2^{-\|\mathsf{B}\|_0}] &\leq \mathbb{E}[2^{-T/2}\mathbb{1}[T/2 \leq \|\mathsf{B}\|_0 \leq T]] + \mathbb{P}(\{\|\mathsf{B}\|_0 < T/2\} \cup \{\|\mathsf{B}\|_0 > T\}) \\ &\leq 2^{-T/2} + 2\exp(-T/36). \end{aligned} \tag{47}$$

Finally, by Eq. (46) and Eq. (47), we obtain

$$\mathrm{H}(U|\tilde{W}_T, \tilde{S}, \mathsf{B}) \leq n(2^{-T/2} + 2\exp(-T/36)). \tag{48}$$

The last step is combining Eq. (44), Eq. (45), and Eq. (48) to lower bound $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ as

$$\begin{aligned}\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= n - \mathrm{H}(U|\tilde{W}_T, \tilde{S}) \\ &\geq n - \left[n2^{-n^2} + n\exp(-n^2/18) + 2n^5 2^n \exp(-2^n/n) + 12n^3\exp(-n^2/18) + 1\right]\end{aligned} \tag{49}$$

Fig. 2 shows the upper bound on $n - \mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ in Eq. (49) as a function of n. As seen for $n \geq 16$, $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq n - 1.1$, and the lower bound on $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ is increasing.

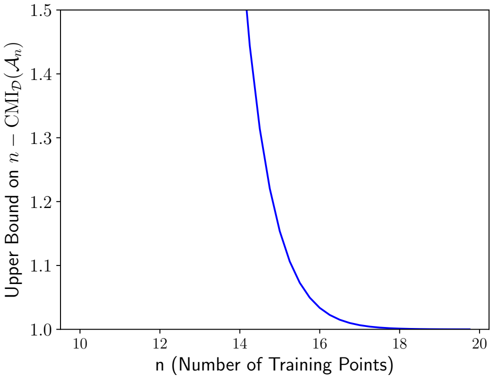

Figure 2: Upper bound on $n - \mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ in Eq. (49).

Hence, we obtain

$$\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \Omega(n),$$

which was to be shown.

Appendix D. Proof of Theorem 9

The construction for proving this theorem is exactly the same as in Theorem 4.

D.1. Lower Bound on the Residual

For the case that $\sigma^2 \leq \mathrm{var}^{\star}_{(n)}$, we showed in Theorem 4 that for sufficently large n, $\mathbb{E}[\Delta_\sigma(W_T) + \hat{\Delta}_\sigma(W_T)] = M_{\text{res}} \in \Omega(1)$. Since the loss function is 4-Lipschitz and the space has radius of 1, we have

$$\Delta_\sigma(W_T) + \hat{\Delta}_\sigma(W_T) \leq 2L\|W_T - \tilde{W}\| \leq 4LR = 16 \quad \text{a.s.}$$

Then, we invoke Lemma 27 with $m = \tilde{m} = 16$ and $a = M_{\text{res}}/2$ to obtain

$$\mathbb{P}(\Delta_\sigma(W_T) + \hat{\Delta}_\sigma(W_T) > M_{\text{res}}/2) \geq \frac{M_{\text{res}}}{32}. \tag{50}$$

D.2. Lower Bound on the Conditional-PAC Bayes Bound

First of all, Lemma 29 implies that $\mathbb{E}^{S_n}[\mathrm{KL}(Q(S_n)\|\frac{1}{2^n}\sum_{u\in\{0,1\}^n} Q(\tilde{S}_u))]$ is bounded by n a.s. In Theorem 4 we showed that given $\sigma^2 \geq \mathrm{var}^{\star}_{(n)}$ for sufficiently large n, we have

$$\mathbb{E}[\mathbb{E}^{S_n}[\mathrm{KL}(Q(S_n)\|\frac{1}{2^n}\sum_{u\in\{0,1\}^n} Q(\tilde{S}_u))]] = \mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq 0.2n.$$

Then, we use Lemma 27, with the following parameters: $\tilde{m} = m = n$ and $a = 0.1n$ to obtain

$$\mathbb{P}\Big(\mathbb{E}^{S_n}[\mathrm{KL}(Q(S_n)\|\frac{1}{2^n}\sum_{u\in\{0,1\}^n} Q(\tilde{S}_u))] > 0.1n\Big) \geq \frac{\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) - 0.1n}{n - 0.1n} \geq \frac{1}{9}. \tag{51}$$

D.3. Lower Bound on the Classical PAC-Bayes Bound

For every $s \in \{0,1\}^{n\times d}$, let $Q(s)$ denote the posterior, and $P = \mathbb{E}[Q(\tilde{S}_n)]$ denote the prior. By construction, the training set S_n takes all the values in $\{0,1\}^{n\times d}$ uniformly at random. Therefore, by Lemma 29, we have

$$\mathrm{KL}(Q(S_n)\|P) = \mathrm{KL}(Q(S_n)\|\frac{1}{2^{nd}}\sum_{s\in\{0,1\}^{n\times d}} Q(s)) \leq nd \quad \text{a.s.}$$

Consider the estimator $\Psi : \mathcal{W} \to \{0,1\}^d$ in Eq. (34). For every $s \in \{0,1\}^d$, let $\hat{Q} : \{0,1\}^{n\times d} \to \mathcal{M}_1(\{0,1\}^d)$ be the pushforward of $Q(s)$ through Ψ. Similarly, we can define $\hat{P} \in \mathcal{M}_1(\{0,1\}^d)$ as the pushforward of P using Ψ.

By the data-processing inequality for the KL divergence [**polyanskiy2014lecture**], we have

$$\mathrm{KL}(Q(S)\|P) \geq \mathrm{KL}(\hat{Q}(S)\|\hat{P}) \quad \text{a.s.} \tag{52}$$

We claim that $\hat{P} = \mathbb{E}[\hat{Q}(S_n)]$. By a slight abuse of notation, for every $b \in \{0,1\}^d$, let $\hat{P}(b)$ denote the probability assigned to b by $\hat{P}$. Also, for every $s \in \{0,1\}^{n\times d}$ and a (measurable) set $A \subseteq \mathcal{W}$ let $Q(s)(A)$ be the measure assigned to set A by $Q(s)$. Similarly, we can define $P(A)$. Equipped with these notations, we can write

$$\begin{aligned}\hat{P}(b) &= \int_{w\in\mathcal{W}} P(dw)\mathbb{1}[\Psi(w)=b]\\ &= \int_{w\in\mathcal{W}} \frac{1}{2^{nd}} \sum_{s\in\{0,1\}^{n\times d}} Q(s)(dw)\mathbb{1}[\Psi(w)=b]\\ &= \frac{1}{2^{nd}} \sum_{s\in\{0,1\}^{n\times d}} \int_{w\in\mathcal{W}} Q(s)(dw)\mathbb{1}[\Psi(w)=b].\end{aligned}$$

Here, the second step is by the definition of the prior, and the last step follows from Fubini's theorem. Notice that the expression in the last step is $\mathbb{E}[\hat{Q}(S_n)]$ as was to be shown.

Recall the definition of the bad coordinates. Define the *bad coordinate profile* of $s\in\{0,1\}^{n\times d}$ as a binary vector of length d such that its $i-$coordinate is one if and only if i is a bad coordinate, and it is zero otherwise. For every $b\in\{0,1\}^d$, define set

$$\mathcal{S}_b = \{s\in\{0,1\}^{n\times d}|\text{bad coordinate profile of } s \text{ is } b\ \}.$$

By construction, each coordinate is a bad coordinate independently with probability 2^{-n}. Therefore

$$\mathbb{P}(S_n\in\mathcal{S}_b) = 2^{-n\|b\|_0}(1-2^{-n})^{d-\|b\|_0}. \tag{53}$$

In what follows, for every $b\in\{0,1\}^d$ that satisfies $T/2\le\|b\|_0\le T$, we provide an upper bound on $\mathrm{KL}(\hat{Q}(s)\|\hat{P})$ given $s\in\mathcal{S}_b$. We can write

$$\begin{aligned}\mathrm{KL}(\hat{Q}(s)\|\hat{P}) &= \mathrm{KL}(\hat{Q}(s)\|\mathbb{E}[\hat{Q}(S_n)])\\ &\le \mathrm{KL}(\hat{Q}(s)\|\mathbb{P}(S_n\in\mathcal{S}_b)\mathbb{E}^{S_n\in\mathcal{S}_b}[\hat{Q}(S_n)]+\mathbb{P}(S_n\notin\mathcal{S}_b)\mathbb{E}^{S_n\notin\mathcal{S}_b}[\hat{Q}(S_n)]).\end{aligned}$$

The last line follows from the law of total expectation. Then, we invoke Lemma 28, to obtain

$$\begin{aligned}&\mathrm{KL}(\hat{Q}(s)\|\mathbb{P}(S_n\in\mathcal{S}_b)\mathbb{E}^{S_n\in\mathcal{S}_b}[\hat{Q}(S_n)]+\mathbb{P}(S_n\notin\mathcal{S}_b)\mathbb{E}^{S_n\notin\mathcal{S}_b}[\hat{Q}(S_n)])\\ &\le -\log(\mathbb{P}(S_n\in\mathcal{S}_b))+\mathrm{KL}(\hat{Q}(s)\|\mathbb{E}^{S_n\in\mathcal{S}_b}[\hat{Q}(S_n)]).\end{aligned}$$

First, we analyze $\log(\mathbb{P}(S_n\in\mathcal{S}_b))$. By Eq. (53), we have

$$-\log(\mathbb{P}(S_n\in\mathcal{S}_b)) = n\|b\|_0+(d-\|b\|_0)\log\left(\frac{1}{1-2^{-n}}\right).$$

Since $T/2\le\|b\|_0\le T$, we have $n\|b\|_0\le nT$. Then, using the inequality $-\log(1-x)\le\frac{x}{1-x}$ for $x\le 1$, we obtain $-\log(1-2^{-n})\le {}^{2^{-n}}/_{1-2^{-n}}$. Therefore,

$$\begin{aligned}(d-\|b\|_0)\log\left(\frac{1}{1-2^{-n}}\right) &\le d\log\left(\frac{1}{1-2^{-n}}\right)\\ &\le \frac{d2^{-n}}{1-2^{-n}}\\ &\le 2d2^{-n}.\end{aligned}$$

Finally setting $d = 0.75T2^n$, we obtain the following upper bound

$$-\log(\mathbb{P}(S_n \in \mathcal{S}_b)) \leq \frac{5}{2}nT. \tag{54}$$

Next, we provide an upper bound on $\mathrm{KL}(\hat{Q}(s)\|\mathbb{E}^{\tilde{S}\in\mathcal{S}_b}[\hat{Q}(\tilde{S})])$. In Eq. (39), we analyzed the error probability of the estimator Ψ conditioned on the training set. In particular, we proved that for every training set whose number of bad coordinates is between $T/2$ and T, we have almost surely

$$\begin{aligned}\mathbb{P}^{S_n}(\exists i \in [d]\ \Psi(W_T+\xi)(i) \neq b(i)) &\leq n^2 2^n \exp(-2^n/n)\\ &\triangleq p_{\text{error}}.\end{aligned}$$

It implies that for all $s \in \mathcal{S}_b$ with $T/2 \leq \|b\|_0 \leq T$, $\hat{Q}(s)(b) \geq 1 - p_{\text{error}}$ and $\sum_{b'\neq b}\hat{Q}(s)(b') \leq p_{\text{error}}$. For notational convenience let $\mathbb{E}^{S_n\in\mathcal{S}_b}[\hat{Q}(S_n)] \triangleq Q_b$. By the definition of the KL divergence, we can write

$$\begin{aligned}\mathrm{KL}(\hat{Q}(s)\|\hat{Q}_b) &= \sum_{b'\in\{0,1\}^d}\hat{Q}(s)(b')\log\Big(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')}\Big)\\ &= \hat{Q}(s)(b)\log\Big(\frac{\hat{Q}(s)(b)}{\hat{Q}_b(b)}\Big) + \sum_{b'\in\{0,1\}^d,b'\neq b}\hat{Q}(s)(b')\log\Big(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')}\Big).\end{aligned} \tag{55}$$

Since for all $s \in \mathcal{S}_b$, $\hat{Q}(s)(b) \geq 1 - p_{\text{error}}$, we have $\hat{Q}_b(b) \geq 1 - p_{\text{error}}$. Therefore, we have

$$\begin{aligned}\hat{Q}(s)(b)\log\Big(\frac{\hat{Q}(s)(b)}{\hat{Q}_b(b)}\Big) &\leq \hat{Q}(s)(b)\log\Big(\frac{\hat{Q}(s)(b)}{1-p_{\text{error}}}\Big)\\ &\leq -\log(1-p_{\text{error}}).\end{aligned} \tag{56}$$

The last step follows from $0 \leq \hat{Q}(s)(b) \leq 1$. Conditioned on $S_n \in \mathcal{S}_b$, the distribution of the training set is uniform over the set $\mathcal{S}_b$. Using this observation, for every $b' \in \{0,1\}^d$, we can write

$$\begin{aligned}\log\Big(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')}\Big) &= \log\Big(\frac{\hat{Q}(s)(b')}{\frac{1}{|\mathcal{S}_b|}\sum_{b'\in\mathcal{S}_b}\hat{Q}(s)(b')}\Big)\\ &\leq \log(|\mathcal{S}_b|)\\ &\leq \log(2^{nd}).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\sum_{b'\in\{0,1\}^d,b'\neq b}\hat{Q}(s)(b')\log\Big(\frac{\hat{Q}(s)(b')}{\hat{Q}_b(b')}\Big) &\leq nd\sum_{b'\in\{0,1\}^d,b'\neq b}\hat{Q}(s)(b')\\ &\leq nd p_{\text{error}}.\end{aligned} \tag{57}$$

By Eq. (54), Eq. (55), Eq. (56), and Eq. (57), we obtain

$$\begin{aligned}-\log(\mathbb{P}(S_n \in \mathcal{S}_b)) + \mathrm{KL}(\hat{Q}(s)\|\mathbb{E}^{\tilde{S}\in\mathcal{S}_b}[\hat{Q}(\tilde{S})]) &\leq \frac{5}{2}nT - \log(1-p_{\text{error}}) + nd p_{\text{error}}\\ &\leq \frac{5}{2}nT + \frac{p_{\text{error}}(nd+1)}{1-p_{\text{error}}}.\end{aligned}$$

Setting the parameters, we can see that $\frac{p_{\text{error}}(nd+1)}{1-p_{\text{error}}} \leq 1$ for $n \geq 8$.

Thus, we obtain that for every $s \in \mathcal{S}_b$ such that $T/2 \leq \|b\|_0 \leq T$, we have

$$\mathrm{KL}(\hat{Q}(s)\|\hat{P}) \leq \frac{5}{2}nT + 1, \tag{58}$$

for $n \geq 8$.

Note that the upper bound in Eq. (58) provides a *uniform* upper bound for every $s \in \mathcal{S}_b$ such that $T/2 \leq \|b\|_0 \leq T$. Therefore, by a simple contraposition we have

$$\begin{aligned}&\{s \in \{0,1\}^{n\times d} | \text{the number of bad coordinates of } s \in \{T/2,\dots,T\}\ \} \\ &\subseteq \{s \in \{0,1\}^{n\times d} | \mathrm{KL}(\hat{Q}(s)\|\hat{P}) \leq \frac{5}{2}nT + 1\}.\end{aligned}$$

By considering the complement of the above statement we obtain

$$\begin{aligned}&\{s \in \{0,1\}^{n\times d} | \mathrm{KL}(\hat{Q}(s)\|\hat{P}) > \frac{5}{2}nT + 1\} \\ &\subseteq \{s \in \{0,1\}^{n\times d} | \text{the number of bad coordinates of } s \notin \{T/2,\dots,T\}\ \}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{P}(\mathrm{KL}(\hat{Q}(S_n)\|\hat{P}) > \frac{5}{2}nT + 1) &\leq 1 - \mathbb{P}(T/2 \leq \|\mathsf{B}\|_0 \leq T) \\ &\leq 2\exp(-T/36).\end{aligned}$$

Here, the line follows from Eq. (15).

Next, we provide a lower bound on $\mathbb{E}[\mathrm{KL}(\hat{Q}(S_n)\|\hat{P})]$. Let random variable B denote the bad coordinate profile of S_n. Notice that $\mathbb{E}[\mathrm{KL}(\hat{Q}(S_n)\|\hat{P})] = I(S_n;\hat{\mathsf{B}})$ where $\hat{\mathsf{B}}$ is the estimate of B using the estimator Ψ. We have $I(\hat{\mathsf{B}};S_n) = \mathrm{H}(S_n) - \mathrm{H}(S_n|\mathsf{B})$. By construction, $\mathrm{H}(S_n) = nd$. Since B is a function of S_n, we have $\mathrm{H}(S_n,\mathsf{B}|\hat{\mathsf{B}}) = \mathrm{H}(S_n|\hat{\mathsf{B}})$. Then, by the chain rule for the entropy we can write $\mathrm{H}(S_n,\mathsf{B}|\hat{\mathsf{B}}) = \mathrm{H}(\mathsf{B}|\hat{\mathsf{B}}) + \mathrm{H}(S_n|\mathsf{B},\hat{\mathsf{B}})$. By conditioning on B, we know the exact values for the bad coordinates in S_n. Therefore, the cardinally of the possible values for each data-point, conditioned on B, cannot be more that $2^{d-\|\mathsf{B}\|_0}$. Therefore, we have $\mathrm{H}(S_n|\mathsf{B},\hat{\mathsf{B}}) \leq n(d - \mathbb{E}[\|\mathsf{B}\|_0])$ which gives us $\mathrm{H}(\mathsf{B}|\hat{\mathsf{B}}) + \mathrm{H}(S_n|\mathsf{B},\hat{\mathsf{B}}) \leq \mathrm{H}(\mathsf{B}|\hat{\mathsf{B}}) + n(d - \mathbb{E}[\|\mathsf{B}\|_0])$. By Fano's inequality in Lemma 24, we have $\mathrm{H}(\mathsf{B}|\hat{\mathsf{B}}) \leq 1 + \mathbb{P}(\hat{\mathsf{B}} \neq \mathsf{B})\mathrm{H}(\mathsf{B})$. Therefore, we obtain

$$\begin{aligned}\mathbb{E}[\mathrm{KL}(\hat{Q}(S_n)\|\hat{P})] &= I(S_n;\hat{\mathsf{B}}) \\ &\geq n\mathbb{E}[\|\mathsf{B}\|_0] - 1 - \mathbb{P}(\hat{\mathsf{B}} \neq \mathsf{B})\mathrm{H}(\mathsf{B}) \\ &\geq nd2^{-n} - (n+1)d2^{-n}\mathbb{P}(\hat{\mathsf{B}} \neq \mathsf{B}) - 1 \\ &\geq 1.5n^3(1 - 2\mathbb{P}(\hat{\mathsf{B}} \neq \mathsf{B})) - 1.\end{aligned}$$

Here, we used the following facts. $\mathbb{E}[\|\mathsf{B}\|_0] = \mathbb{E}[\sum_{i=1}^d \mathsf{B}(i)] = d\mathbb{E}[\mathsf{B}(1)] = d2^{-n}$ since each element of S_n is i.i.d.and each column is a bad coordinate with probability 2^{-n}. Also, with the similar reasoning we obtain $\mathrm{H}(\mathsf{B}) = \mathrm{H}(\mathsf{B}(1),\dots,\mathsf{B}(d)) = d\mathrm{H}_b(2^{-n})$, where for $x \in [0,1]$ $\mathrm{H}_b(x) = -x\log(x) - (1-x)\log(1-x)$ is the binary entropy function. Also, we have used the the well-known inequality $\mathrm{H}_b(x) \leq -x\log(x) + x$. Then, our analysis of the error probability of the estimator Ψ in Appendix C.4 implies that for $n \geq 10$, the following lower bound holds:

$$\mathbb{E}[\mathrm{KL}(\hat{Q}(S_n)\|\hat{P})] \geq 1.2n^3 - 1.$$

In the next step, we invoke Lemma 27 with the following parameters $\hat{m} = nd$, $m = \frac{5}{2}nT + 1 = 5n^3 + 1$, and $a = 0.6n^3 - 0.5$ to write

$$\begin{aligned}
&\mathbb{P}(\mathrm{KL}(\hat{Q}(S_n)\|\hat{P}) \geq 0.6n^3 - 0.5) \\
&\geq \frac{\mathbb{E}[\mathrm{KL}(\hat{Q}(S_n)\|\hat{P})] - a - (nd - (\frac{5}{2}nT + 1))\mathbb{P}(X \geq \frac{5}{2}nT + 1)}{5n^3 + 1 - a} \\
&\geq \frac{0.6n^3 - 0.5 - 3n^3 2^n \exp(-n^2/18)}{4.4n^3 + 1.5}.
\end{aligned}$$

By numerical evaluations, we can see that the lower bound is greater than 0.1 for $n \geq 16$.

From Eq. (52), we have

$$\begin{aligned}
\mathbb{P}(\mathrm{KL}(Q(S_n)\|P) > 0.6n^3 - 0.5) &\geq \mathbb{P}(\mathrm{KL}(\hat{Q}(S_n)\|\hat{P}) \geq 0.6n^3 - 0.5) \\
&\geq 0.1, \qquad (59)
\end{aligned}$$

for $n \geq 16$ as was to be shown.

D.4. Concluding the Proof

In summary, in Eq. (50), Eq. (51), and Eq. (59), we have shown there exist constants $\alpha_1 \in \mathbb{R}_+$, $\alpha_2 \in \mathbb{R}_+$, $\alpha_3 \in \mathbb{R}_+$, $\beta_1 \in (0,1)$, and $\beta_2 \in (0,1)$ such that for sufficiently large n,

1. $\mathbb{P}\Big(\Delta_\sigma(W_T) + \hat{\Delta}_\sigma(W_T) > \alpha_1 \text{ or } \frac{\mathbb{E}^{S_n}[\mathrm{KL}(Q(S_n)\|\frac{1}{2^n}\sum_{u\in\{0,1\}^n} Q(\tilde{S}_u))]}{n} > \alpha_2\Big) \geq 1 - \beta_1.$
2. $\mathbb{P}\Big(\Delta_\sigma(W_T) + \hat{\Delta}_\sigma(W_T) > \alpha_1 \text{ or } \frac{\mathrm{KL}(Q(S_n)\|\mathbb{E}[Q(S_n)])}{n} > \alpha_3\Big) \geq 1 - \beta_2.$

For notational convenience, let $\mathrm{Bad\ Event}_1$ and $\mathrm{Bad\ Event}_2$ denote the first and second event above.

Next, we show how this result implies the failure of PAC-Bayes bounds. Consider the decomposition of the generalization error of GD with respect to the surrogate

$$\mathbb{E}^{S_n}\Big[\mathrm{F}_{\mathcal{D}}(W_T) - \hat{\mathrm{F}}_{S_n}(W_T)\Big] \leq \mathbb{E}^{S_n}\Big[\mathrm{F}_{\mathcal{D}}(\tilde{W}) - \hat{\mathrm{F}}_{S_n}(\tilde{W})\Big] + \hat{\Delta}_\sigma(W_T) + \Delta_\sigma(W_T),$$

Let $\mathsf{complexity}(n)$ denote both $C_{\text{clas}}(n) \triangleq \mathrm{KL}(Q(S_n)\|\mathbb{E}[Q(S_n)])$ and $C_{\text{cond}}(n) \triangleq \mathbb{E}^{S_n}[\mathrm{KL}(Q(S_n)\|\frac{1}{2^n}\sum_{u\in\{0,1\}^n} Q(\tilde{S}_u))]$. Let $\delta < 1 - \max\{\beta_1, \beta_2\}$. Assume we instantiate the PAC-Bayes bounds with the confidence of $1-\delta$. Then, by a simple application of the union bound we have

$$\mathbb{P}\Bigg(\Bigg\{\mathbb{E}^{S_n}\Big[\mathrm{F}_{\mathcal{D}}(\tilde{W}) - \hat{\mathrm{F}}_{S_n}(\tilde{W})\Big] \in \mathcal{O}\Bigg(LR\sqrt{\frac{C_{\text{clas}}(n) + \log(n/\delta)}{n}}\Bigg)\Bigg\} \text{ and Bad Event}_1\Bigg) \geq 1 - \delta - \beta_1,$$

$$\mathbb{P}\Bigg(\Bigg\{\mathbb{E}^{S_n}\Big[\mathrm{F}_{\mathcal{D}}(\tilde{W}) - \hat{\mathrm{F}}_{S_n}(\tilde{W})\Big] \in \mathcal{O}\Bigg(LR\sqrt{\frac{C_{\text{cond}}(n) + \log(n/\delta)}{n}}\Bigg)\Bigg\} \text{ and Bad Event}_2\Bigg) \geq 1 - \delta - \beta_2.$$

Thus, we conclude that with probability at least $1 - \delta - \max\{\beta_1, \beta_2\}$ (over the randomness in the training set) for every σ we have

$$\max\{LR\sqrt{\frac{\mathsf{complexity}(n) + \log(n/\delta)}{n}}, \hat{\Delta}_\sigma(W_T) + \Delta_\sigma(W_T)\} \in \Omega(1),$$

as was to be shown.

Appendix E. Proof of Theorem 17

Let $d \in \mathbb{N}$ and $\mathcal{Z} = \{\mathsf{e}(i) : i \in d\}$, that is, the set of all coordinate vectors in $\{0,1\}^d$, where

$$\mathsf{e}(i) = (\underbrace{0, \ldots, 0}_{i-1 \text{ times}}, 1, \underbrace{0, \ldots, 0}_{d-i \text{ times}}).$$

Let the data distribution on the input be the uniform distribution, that is $\mathcal{D} = \mathsf{Uniform}(\mathcal{Z})$. Then, we consider the simple convex, 1-Lipschitz loss function $f(w, z) = -\langle w, z\rangle$. Moreover, we consider that the weights w are in a unit ball on $\mathbb{R}^d$, that is $\mathcal{W} = \{w : \|w\| \leq 1\}$. Therefore, the problem is in the CLB class.

Next, we analyze the dynamics of GD. The empirical loss is given by

$$\hat{\mathrm{F}}_{S_n}(w) = -\langle w, \hat{\mu}\rangle,$$

where $\hat{\mu}$ is the empirical mean of the instances in the training set, i.e., $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n Z_i$. Also, we have that $\partial\hat{\mathrm{F}}_{S_n}(w) = -\hat{\mu}$ for all $w \in \mathcal{W}$. Considering the update rule of GD, i.e. $W_{t+1} = \Pi_{\mathcal{W}}(W_t + \eta\hat{\mu})$, one can show by induction that

$$W_t = \begin{cases} \eta t\hat{\mu} & \eta t\|\hat{\mu}\| \leq 1 \\ \frac{\hat{\mu}}{\|\hat{\mu}\|} & \text{Otherwise} \end{cases}. \tag{60}$$

Now consider the $\tilde{S}$-measurable random variable E that is equal to one if and only if all the data instances in the supersample are distinct. That is

$$E = \mathbb{1}[\tilde{Z}_{u,i} \neq \tilde{Z}_{v,j} \text{ for all } i, j \in [n] \text{ and all } u, v \in \{0,1\}].$$

As in the *birthday paradox problem* [**mitzenmacher2017probability**], we may bound the probability that $E = 1$ as follows

$$\begin{aligned}\mathbb{P}(E=1) &= \prod_{k=0}^{2n-1}\left(1-\frac{k}{d}\right) \\ &\geq \left(1-\frac{2n-1}{d}\right)^{2n-1}, \qquad (61)\end{aligned}$$

This way, we may engineer a dimension d for which $\mathbb{P}(E=1) \geq c$ for all $n \geq 1$, where c is a constant probability, independent of n. Solving for Eq. (61) results in

$$d \geq \frac{2n-1}{1-c^{1/(2n-1)}}.$$

For instance, for $c = 0.1$, a dimension $d = 2n^2$ suffices, and therefore $\mathbb{P}(E = 0) \leq 0.9$. Now, we are ready to study what happens to both the individual conditional mutual information $I(W_T; U_i|\tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ and the evaluated mutual information $e\mathrm{CMI}_{\mathcal{D}}(f(\mathrm{GD}_n))$ in this particular setting.

E.1. Individual conditional mutual information

Note that the individual CMI may be written as follows

$$\begin{aligned}I(W_T; U_i|\tilde{Z}_{0,i}, \tilde{Z}_{1,i}) &= \mathrm{H}(U_i|\tilde{Z}_{0,i}, \tilde{Z}_{1,i}) - \mathrm{H}(U|W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) \\ &= \mathrm{H}(U_i) - \mathrm{H}(U_i|W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) \\ &= \log 2 - \mathrm{H}(U_i|W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}), \qquad (62)\end{aligned}$$

where the second and third equations follow from $U_i \perp (\tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ and $H(U_i) = \log 2$, respectively.

From Eq. (60), we can see that if we know that the non-zero coordinates of W_T are precisely the coordinates of the training samples. That is, if $\tilde{Z}_{u,i} = \mathsf{e}(k)$, then $W_T(k) \neq 0$. Therefore, under the event $E = 1$, one can precisely determine if sample $\tilde{Z}_{0,i}$ or sample $\tilde{Z}_{1,i}$ was used for training after observing W_T. That is, one can completely determine U_i from $(W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ or, equivalently, $\mathbb{E}\Big[\mathrm{H}^{W_T,\tilde{Z}_{0,i},\tilde{Z}_{1,i},E}(U_i)\mathbb{1}[E=1]\Big] = 0$. We may use this fact to bound $\mathrm{H}(U_i|W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i})$ and obtain the desired result. Namely,

$$\begin{aligned}\mathrm{H}(U_i|W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}) &= \mathrm{H}(U_i|W_T, \tilde{Z}_{0,i}, \tilde{Z}_{1,i}, E) \\ &= \mathbb{E}\Big[\mathrm{H}^{W_T,\tilde{Z}_{0,i},\tilde{Z}_{1,i},E}(U_i)\mathbb{1}[E=0]\Big] + \mathbb{E}\Big[\mathrm{H}^{W_T,\tilde{Z}_{0,i},\tilde{Z}_{1,i},E}(U_i)\mathbb{1}[E=1]\Big] \\ &\leq 0.9\log 2, \qquad (63)\end{aligned}$$

where the last inequality follows from upper bounding $\Big[\mathrm{H}^{W_T,\tilde{Z}_{1,i},\tilde{Z}_{2,i},E}(U_i)\Big]$ by $\log 2$ and the facts that $\mathbb{E}\Big[\mathrm{H}^{W_T,\tilde{Z}_{0,i},\tilde{Z}_{1,i},E}(U_i)\mathbb{1}[E=1]\Big] = 0$ and $\mathbb{P}(E=0) \leq 0.9$.

Finally, combining Eq. (62) and Eq. (63) results in

$$I(W_T; U_i|\tilde{Z}_{0,i}, \tilde{Z}_{1,i}) \geq \log 2 - 0.9\log 2 \in \Omega(1),$$

and completes the proof.

E.2. Evaluated conditional mutual information

Note that the evaluated CMI may be written as follows

$$\begin{aligned} e\mathrm{CMI}_{\mathcal{D}}(f(\mathrm{GD}_n)) &= I(F; U|\tilde{S}) \\ &= \mathrm{H}(U|\tilde{S}) - \mathrm{H}(U|\tilde{S}, F) \\ &= \mathrm{H}(U) - \mathrm{H}(U|F, \tilde{S}) \\ &= n \log 2 - \mathrm{H}(U|F, \tilde{S}), \end{aligned} \tag{64}$$

where the third and fourth equations follow from $U \perp \tilde{S}$ and $H(U) = n \log 2$, respectively.

Then, as in the previous subsection, the proof relies in the fact that U can be completely determined by the loss vector F under the event $E = 1$. More precisely, note that $F_{u,i} = f(W_T, \tilde{Z}_{u,i}) = -\langle W_T, \tilde{Z}_{u,i} \rangle$. Also, remember from the previous subsection that the non-zero coordinates of W_T are precisely the non-zero coordinates of the samples that are used for training. Therefore, under the event $E = 1$, $F_{u,i} = 0$ if and only if $\tilde{Z}_{u,i}$ was not used for training and therefore $Z_i = \tilde{Z}_{1-u,i}$. Hence, one can completely determine U from F or, equivalently, $\mathbb{E}\Big[\mathrm{H}^{F,\tilde{S},E}(U)\mathbb{1}[E = 1]\Big] = 0$. We may use this fact to bound $\mathrm{H}(U|F, \tilde{S})$ and obtain the desired result. Namely,

$$\begin{aligned} \mathrm{H}(U|F, \tilde{S}) &= \mathrm{H}(U|F, \tilde{S}, E) \\ &= \mathbb{E}\Big[\mathrm{H}^{F,\tilde{S},E}(U)\mathbb{1}[E = 1]\Big] + \mathbb{E}\Big[\mathrm{H}^{F,\tilde{S},E}(U)\mathbb{1}[E = 0]\Big], \\ &\leq n \cdot 0.9 \log 2 \end{aligned} \tag{65}$$

where the first line follows since E is $\tilde{S}$-measurable, and the last inequality follows from upper bounding $\mathrm{H}^{\tilde{S},F,G}(U)$ by $n \log 2$ and the facts that $\mathbb{E}\Big[\mathrm{H}^{F,\tilde{S},E}(U)\mathbb{1}[E = 1]\Big] = 0$ and $\mathbb{P}(E = 0) \leq 0.9$.

Finally, combining Eq. (64) and Eq. (65) results in

$$e\mathrm{CMI}_{\mathcal{D}}(f(\mathrm{GD}_n)) \geq n \log 2 - n \cdot 0.9 \log 2 \in \Omega(n),$$

and completes the proof.

Appendix F. Helper Lemmata

Lemma 21 ([vershynin2018high]) *Let $X \sim \mathcal{N}(0, \mathbb{I}_d)$. Let us represent $X = R\theta$ where $R = \|X\|$ and $\theta = X/\|X\|$. Then, R and θ are independent random variables. Also, θ is uniformly distributed on the Euclidean sphere $\mathrm{S}^{(\mathrm{d}-1)}$ with the center at the origin.*

Lemma 22 ([laurent2000adaptive]) *Consider random vector $X \sim \mathcal{N}(0, \mathbb{I}_d)$. Then,*

$$\mathbb{P}\bigg(\sum_{i=1}^{d} a(i)X(i)^2 \geq \|a\|_1 + 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\bigg) \leq \exp(-t) \text{ and}$$

$$\mathbb{P}\bigg(\sum_{i=1}^{d} a(i)X(i)^2 \geq \|a\|_1 - 2\|a\|_2\sqrt{t}\bigg) \leq \exp(-t)$$

Corollary 23 *Let $\sigma \in \mathbb{R}$, $\delta \in (0,1)$, $d \in \mathbb{N}$, and $d \geq \log \frac{2}{\delta}$. Consider $X \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$, then*

$$\mathbb{P}\Big(d\sigma^2(1 - 2\sqrt{\frac{\log(2/\delta)}{d}}) \leq \|X\|^2 \leq d\sigma^2(1 + 4\sqrt{\frac{\log(2/\delta)}{d}})\Big) \geq 1 - \delta,$$
$$\mathbb{P}\big(\|X\| \leq \sqrt{(1-\alpha)d\sigma^2}\big) \leq 2\exp\Big(-\frac{d\alpha^2}{4}\Big) \textit{ for } \alpha \in [0,1], \textit{ and}$$
$$\mathbb{P}(\|X\| \geq \sqrt{(1+\beta)d\sigma^2}) \leq 2\exp\Big(-\frac{d\beta^2}{16}\Big) \textit{ for } \beta \geq 0.$$

Lemma 24 ([cover2012elements]) *Let X and Y be discrete random variables. Then*

$$\mathrm{H}(X|Y) \leq \mathrm{H}_b(\mathsf{P_e}) + \mathsf{P_e}\mathrm{H}(X) \leq 1 + \mathsf{P_e}\mathrm{H}(X),$$

*where $\mathsf{P_e} = \mathbb{P}(\Psi(Y) \neq X)$ for any (possibly randomized) estimator Ψ of X using Y (See also [**fano1952class**]).*

Lemma 25 *Let $d \in \mathbb{N}_+$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be defined as $g(x) = \max\{\max_{i\in[d]}\{x(i)\}, 0\}$. Then, g is $1-$Lipschitz.*

Proof Let $x \in \mathbb{R}^d$ and $\Delta \in \mathbb{R}^d$. Let $\arg\max_{i\in[d]} \{x(i) + \Delta(i)\} = i^\star$ and $\arg\max_{i\in[d]} \{x(i)\} = j^\star$ (break ties arbitrary). Then,

$$g(x+\Delta) - g(x) = \begin{cases} -x(j^\star) \leq 0 & x(i^\star) + \Delta(i^\star) \leq 0 \text{ and } x(j^\star) > 0 \\ x(i^\star) + \Delta(i^\star) - x(j^\star) < \Delta(i^\star) & x(i^\star) + \Delta(i^\star) > 0 \text{ and } x(j^\star) > 0 \\ 0 & x(i^\star) + \Delta(i^\star) \leq 0 \text{ and } x(j^\star) \leq 0 \\ x(i^\star) + \Delta(i^\star) \leq \Delta(i^\star) & x(i^\star) + \Delta(i^\star) > 0 \text{ and } x(j^\star) \leq 0 \end{cases}$$

The last case follows because $x(i^\star) \leq x(j^\star) \leq 0$, therefore, $x(i^\star) + \Delta(i^\star) \leq \Delta(i^\star)$. Thus, $|g(x+\Delta) - g(x)| \leq \|\Delta\|$, as was to be shown. ■

Lemma 26 *Let f be a convex and $L-$Lipschitz loss function, and $\mathcal{W}$ be a convex and compact domain space with bounded diameter R. Let $\{w_t\}_{t\in[T]}$ denote the output of GD algorithm with a constant step size η. Then, we have*

$$f(w_T) - \min_{w\in\mathcal{W}} f(w) \leq \frac{R^2}{2\eta T} + \frac{(\log(T)+2)\eta L^2}{2}$$

Proof Let $g_t \in \partial f(w_t)$. From [**lastiterate**],

$$f(w_T) - \min_{w\in\mathcal{W}} f(w) \leq \frac{1}{T}\sum_{t=1}^{T}(f(w_t) - \min_{w\in\mathcal{W}} f(w)) + \frac{1}{2}\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\eta\|g_t\|^2.$$

Since $\|g_t\| \leq L$, the second term can be upper bounded by $\frac{\eta L^2}{2}\sum_{k=1}^{T-1}\frac{1}{k}$. Then, by the well-known bounds on the Harmonic numbers we have $\frac{\eta L^2}{2}\sum_{k=1}^{T-1}\frac{1}{k} \leq \frac{\eta L^2}{2}(\log(T-1)+1) \leq \frac{\eta L^2}{2}(\log(T)+1)$. For the first term, from [**bubeck2015convex**], we have $\frac{1}{T}\sum_{t=1}^{T}(f(w_t) - \min_{w\in\mathcal{W}} f(w)) \leq \frac{R^2}{2\eta T} + \frac{\eta L^2}{2}$. Combining these two upper bounds proves the lemma. ■

Lemma 27 *Let X be a random variable, $\tilde{m} \geq 0$ be a constant such that $0 \leq X \leq \tilde{m}$ a.s. Let $m \in \mathbb{R}$ be such that $0 < m \leq \tilde{m}$. Then, for every $0 \leq a < m$, we have*

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

Proof The following holds almost surely:

$$X \leq a\mathbb{1}[X \leq a] + m\mathbb{1}[a < X \leq m] + \tilde{m}\mathbb{1}[m < X].$$

Taking an expectation concludes the proof. ■

Lemma 28 ([9531956]) *Let $M \in \mathbb{N}$ and $\mathcal{Y}$ be a measurable space. Let also $P \in \mathcal{M}_1(\mathcal{Y})$ and $Q_i \in \mathcal{M}_1(Y)$ for all $i \in [M]$ be probability measures. If $\alpha_i \in (0,1)$ such that $\sum_{i=1}^M \alpha_i = 1$,*

$$\mathrm{KL}(P \| \sum_{i=1}^{M} \alpha_i Q_i) \leq \min_{i \in [M]} \Big\{ \mathrm{KL}(P \| Q_i) - \log(\alpha_i) \Big\}.$$

Lemma 29 *Let $\mathcal{Y}$ be a measurable space. Let $M \in \mathbb{N}$ and $P_i \in \mathcal{M}_1(\mathcal{Y})$ for $i \in [M]$ be M probability measures. Then, for every $i \in [M]$, we have*

$$\mathrm{KL}(P_i \| \sum_{j=1}^{M} \frac{1}{M} P_j) \leq \log(M).$$

Proof A direct application of Lemma 28 gives us the result. ■