Exactly Tight Information-Theoretic Generalization Error Bound for the Quadratic Gaussian Problem

Ruida Zhou, *Member, IEEE*, Chao Tian, *Senior Member, IEEE*, and Tie Liu, *Senior Member, IEEE*

Abstract—**We provide a new information-theoretic generalization error bound that is exactly tight (i.e., matching even the constant) for the canonical quadratic Gaussian (location) problem. Most existing bounds are order-wise loose in this setting, which has raised concerns about the fundamental capability of information-theoretic bounds in reasoning the generalization behavior for machine learning. The proposed new bound adopts the individual-sample-based approach proposed by Bu et al., but also has several key new ingredients. Firstly, instead of applying the change of measure inequality on the loss function, we apply it to the generalization error function itself; secondly, the bound is derived in a conditional manner; lastly, a reference distribution is introduced. The combination of these components produces a KL-divergence-based generalization error bound. We show that although the latter two new ingredients can help make the bound exactly tight, removing them does not significantly degrade the bound, leading to an asymptotically tight mutual-information-based bound. We further consider the vector Gaussian setting, where a direct application of the proposed bound again does not lead to tight bounds except in special cases. A refined bound is then proposed by a decomposition of loss functions, leading to a tight bound for the vector setting.**

Index Terms—**Information theory, machine learning.**

I. Introduction

UNDERSTANDING the generalization behavior and bounding the generalization error of learning algorithms are important subjects of study in machine learning theory. Recently, information-theoretic approaches to bound generalization errors have drawn considerable attention in both the information theory community and the machine learning community [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]. These bounds can provide intuitions by relating to information-theoretic quantities, leading to novel reasoning and revealing deep connections to existing results such as the classic VC-dimension and Rademacher complexity [27]. Information-theoretic bounds can take into account both data distribution and the dependence between data and algorithm output, which cannot be fully captured by the conventional complexity-based bounds.

In classic information theory research, the study of complex communication systems usually starts from the simplest canonical settings. Particularly, the canonical quadratic Gaussian settings have played tremendous roles in the study of both channel coding and source coding [28]. The study of Additive White Gaussian Noise (AWGN) channel under the average power constraint can be traced back to the original paper by Shannon [29] and led to many subsequent developments in wireless communications [30]. Similarly, the Gaussian source compression under the quadratic distortion measure has been studied extensively [31], [32], which led to many well-used designs of data compression and quantization methods. The motivation to study the Gaussian settings can perhaps be explained as follows. Mathematically, the simplicity of the Gaussian settings, the statistic properties of Gaussian distributions (e.g., the central limit theorem guarantees that aggregation of small independent noises will lead approximately to a Gaussian distribution), the optimality of linear estimators, and the connection to information measures (e.g., differential entropy and entropy power inequality) allow the derivation of precise results and exact tight bounds, which can serve as a running ramp for more complex settings. Practically, Gaussian noises and Gaussian sources can be good approximations to random quantities encountered in many applications, further strengthening the motivation to study the Gaussian settings.

In sharp contrast to the classical information theory research, in the study of generalization error bounds, although various more sophisticated settings such as meta-learning [7], [22] and iterative stochastic algorithms [4], [15] have been considered, our understanding of the canonical quadratic Gaussian problem is in fact quite limited. In this problem setting, independent Gaussian samples are observed, and the learning algorithm chooses the sample average as the hypothesis parameter to locate the mean value. The loss function is the squared difference between the samples and this hypothesis parameter. It turns out that earlier information-theoretic bounds are either vacuous [2] or order-wise loose [9], [10], [12], [14]. The only approaches that provide order-wise tight bounds in this setting either only hold asymptotically [13], or have a loose constant and require a careful construction of certain auxiliary probability structure [16].

Manuscript received 24 October 2023; revised 21 January 2024; accepted 19 March 2024. Date of publication 22 March 2024; date of current version 11 April 2024. The work of Chao Tian and Tie Liu was supported in part by the National Science Foundation under Grant DMS-2312173. This work was presented in part at 2023 IEEE International Symposium on Information Theory, Taipei, July 2023 [DOI: 10.1109/ISIT54713.2023.10206951]. *(Corresponding author: Chao Tian.)*

Ruida Zhou was with the Department of Electrical and Computer Engineering, Texas A&M University at College Station, College Station, TX 77843 USA. He is now with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095 USA.

Chao Tian and Tie Liu are with the Department of Electrical and Computer Engineering, Texas A&M University at College Station, College Station, TX 77843 USA (e-mail: chao.tian@tamu.edu).

Digital Object Identifier 10.1109/JSAIT.2024.3380598

In this work, we provide a new information-theoretic bound that is exactly tight (i.e., matching even the constant) for the canonical quadratic Gaussian (location) problem. The proposed new bound adopts the individual-sample-based approach proposed by Bu et al. [9], but also has several key new ingredients. Firstly, instead of applying the change of measure inequality on the loss function, we apply it to the generalization error function itself; secondly, the bound is derived in a conditional manner; lastly, a reference distribution, which bears a certain similarity to the prior distribution in the Bayesian setting, is introduced. The combination of these components produces a general KL-divergence-based generalization error bound. We also show that although the conditional bounding and the reference distribution can make the bound exactly tight, removing them does not significantly degrade the bound, which results in a mutual-information-based bound that is also asymptotically tight in this setting.

In order to further understand the proposed generalization error bound, we consider the vector version of the Gaussian location problem. The samples here are independent Gaussian vectors, and the algorithm is again the sample mean, but the loss function is a general squared matrix norm. We show that a direct application of the proposed bound is no longer tight in this setting except in certain special cases. However, a refined information-theoretic bound that takes advantage of the decomposition of the matrix norm can indeed lead to a tight bound.

The rest of the paper is organized as follows. In Section II we provide the preliminaries and some relevant previous results. The new generalization error bound is provided in Section III, and then applied on the canonical quadratic Gaussian problem in Section IV. The generalized vector setting is considered in Section V. Finally, Section VI concludes the paper, and a few technical proofs are included in the Appendix.

II. Preliminaries

A. Generalization Error

Denote the data domain as $\mathcal{Z}$, e.g., in the supervised learning setting $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature domain and $\mathcal{Y}$ is the label set. The parametric hypothesis class is denoted as $\mathcal{H}_{\mathcal{W}} = \{h_W : W \in \mathcal{W}\}$, where $\mathcal{W}$ is the parameter space. During training, the learning algorithm has access to a sequence of training samples $Z_{[n]} = (Z_1, Z_2, \dots, Z_n)$, where each Z_i is drawn independently from $\mathcal{Z}$ following some unknown probability distribution ξ. The learner can be represented by $P_{W|Z_{[n]}}$, which is a kernel (channel) that (potentially randomly) maps $\mathcal{Z}^n$ to $\mathcal{W}$.

The learner wishes to choose a hypothesis $w \in \mathcal{W}$ to minimize the following population loss, under a given loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$,

$$L_\xi(w) = \mathbb{E}_{\tilde{Z}\sim\xi}\big[\ell(w, \tilde{Z})\big]. \tag{1}$$

The empirical loss of w is

$$L_{Z_{[n]}}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i). \tag{2}$$

The expected generalization error of the learner $P_{W|Z_{[n]}}$ is

$$\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) \triangleq \mathbb{E}_P\big[L_\xi(W) - L_{Z_{[n]}}(W)\big], \tag{3}$$

where the expectation is taken over the distribution $P_{W,Z_{[n]}}$ as the joint distribution implied by the kernel $P_{W|Z_{n]}}$ and the marginal $P_{Z_{[n]}} = \xi^n$.

Assume another distribution $Q_{W,Z_{[n]}}$, where W and $Z_{[n]}$ are independent and the marginal $Q_{Z_{[n]}}$ is the same as $P_{Z_{[n]}}$, i.e., $Q_{W,Z_{[n]}} = Q_W Q_{Z_{[n]}} = Q_W P_{Z_{[n]}}$. The marginal distribution Q_W can be viewed as a prior distribution in this case.[1] For such Q's, apparently, we have

$$\mathrm{gen}\big(\xi, Q_{W|Z_{[n]}}\big) \triangleq \mathbb{E}_Q\big[L_\xi(W) - L_{Z_{[n]}}(W)\big] = 0, \tag{4}$$

where the equality is because $Q_{W,Z_{[n]}} = Q_W P_{Z_{[n]}}$. In other words, when the algorithm does not learn from the data, its generalization error is zero. This fact will be used during the bounding of generalization error, when we apply the change of measure inequality to the generalization error function itself, instead of on the loss function.

B. Variational Representation of the KL Divergence

The Donsker-Varadhan variational representation of KL divergence for a random scalar-valued random function $F = f(X)$ on a random variable X is given by

$$\mathrm{KL}(P||Q) = \sup_f \big\{\lambda \mathbb{E}_P[F] - \ln \mathbb{E}_Q\big[e^{\lambda F}\big]\big\}, \tag{5}$$

where the equality is achieved when $\lambda F^* = \ln \frac{dP}{dQ} + C$ and $\frac{dP}{dQ}$ is the Radon-Nikodym derivative, or in the inequality form

$$\lambda \mathbb{E}_P[F] \leq \mathrm{KL}(P||Q) + \ln \mathbb{E}_Q\big[e^{\lambda F}\big], \quad \forall \lambda \in \mathbb{R}. \tag{6}$$

This inequality is sometimes also referred to as the change of measure inequality [33]. P and Q can be the distributions of the underlying random variable X, or more directly, the distributions of F. In the context of bounding generalization error, examples are $F = \ell(W, Z)$ or $F = L_\xi(W) - \ell(W, Z)$. We remark here that in the variational representation (5), the supremum is taken over the functions f, whereas when we apply the change of measure inequality (6), the function f is usually already fixed, but the distribution Q can be optimized to make the bound tighter.

The centered cumulant generating function of a random variable F is

$$\Lambda_{F,Q}(\lambda) = \ln \mathbb{E}_Q\big[e^{\lambda F}\big] - \lambda \mathbb{E}_Q[F]. \tag{7}$$

Combining it with the inequality above gives

$$\mathrm{KL}(P||Q) + \Lambda_{F,Q}(\lambda) \geq \lambda \mathbb{E}_P[F] - \lambda \mathbb{E}_Q[F], \quad \lambda \in \mathbb{R}. \tag{8}$$

Now if we choose $F = f(W, Z)$, then for any $Z = z$ the conditional version of the above inequality is

$$\begin{aligned}\mathrm{KL}\big(P_{W|Z=z}||Q_{W|Z=z}\big) &+ \Lambda_{F|Z=z, Q_{W|Z=z}}(\lambda) \\ &\geq \lambda \mathbb{E}_P[F|Z=z] - \lambda \mathbb{E}_Q[F|Z=z], \quad \lambda \in \mathbb{R},\end{aligned} \tag{9}$$

[1] In the Bayesian setting, the distribution P is usually used to denote the prior distribution and Q as the posterior (data dependent) distribution. This is reversed from ours, which follows the convention in information-theoretic literature.

where

$$\Lambda_{F|Z=z,Q_{W|Z=z}}(\lambda) = \ln \mathbb{E}_{Q_{W|Z=z}}\left[e^{\lambda F}|Z=z\right] - \lambda \mathbb{E}_{Q_{W|Z=z}}[F|Z=z]. \quad (10)$$

We will simply replace $Z = z$ in the condition with Z when the exact conditional value realization is not specified.

With a positive λ, we obtain

$$\mathbb{E}_P[F] - \mathbb{E}_Q[F] \leq \inf_{\lambda>0}\left\{\frac{\mathrm{KL}(P||Q) + \Lambda_{F,Q}(\lambda)}{\lambda}\right\}, \quad (11)$$

where equality is achieved if and only if

$$\ln\frac{\mathrm{d}P}{\mathrm{d}Q} \in \{\lambda F + b : \lambda \in \mathbb{R}_+,\ b \in \mathbb{R}\}. \quad (12)$$

The equality condition can also be interpreted as requiring us to choose $\mathrm{d}Q \propto \exp(-\lambda F)\ \mathrm{d}P$. When P is the joint distribution of underlying random variables, and Q is the product distribution of their marginals, then $\mathrm{KL}(P||Q)$ reduces to a mutual information term.

To be consistent with past results in the literature, we will sometimes use the following definition. The Legendre dual function on the interval $[0, b)$ for some $0 < b \leq \infty$ is

$$\Lambda^*(x) \triangleq \sup_{\lambda\in[0,b)} (\lambda x - \Lambda(\lambda)). \quad (13)$$

$\Lambda(\lambda)$ is convex and $\Lambda(0) = \Lambda'(0) = 0$. It can be shown that the inverse dual function is

$$\Lambda^{*-1}(y) = \inf_{\lambda\in[0,b)}\left(\frac{y + \Lambda(\lambda)}{\lambda}\right). \quad (14)$$

C. The Scalar Quadratic Gaussian Location Problem

In the Gaussian location problem introduced by Bu et al. [9], data samples are $Z_1, Z_2, \ldots, Z_n \overset{i.i.d.}{\sim} \xi = \mathcal{N}(\mu, \sigma^2)$ and the sample-average algorithm chooses the following hypothesis $W = \frac{1}{n}\sum_{i=1}^n Z_i$. The loss function is the quadratic function given as $\ell(w, z_i) = (w-z_i)^2$, and the sample-average algorithm is in fact empirical risk minimization for this quadratic loss. Note that this problem setting is not a traditional supervised learning setting, therefore, the data point Z does not decompose into the feature X and the label Y. The expected generalization error is

$$\begin{aligned} \mathrm{gen}(\xi, P_{W|Z_{[n]}}) &= \mathbb{E}\left[\left(\tilde{Z} - W\right)^2 - \frac{1}{n}\sum_{i=1}^n (Z_i - W)^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left[\left(\tilde{Z}_i - W\right)^2 - (Z_i - W)^2\right]\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\tilde{Z}_i^2 - Z_i^2 + 2\left(Z_i - \tilde{Z}_i\right)W\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\sigma^2 + \mu^2 - Z_i^2 + 2(Z_i - \mu)W\right), \end{aligned} \quad (15)$$

where $\tilde{Z}_{[n]}$ are n i.i.d. testing samples, independent of everything else, and the expectation is with respect to distribution $P_{\tilde{Z}}P_{Z^n,W}$, where the joint distribution $P_{Z^n,W}$ is induced by the algorithm $W = \frac{1}{n}\sum_{i=1}^n Z_i$. It is straightforward to show that the true generalization error is in fact $2\sigma^2/n$.

In this work, we shall consider a slightly more general version of the sample-average algorithm that $W = \sum_{i=1}^n \alpha_i Z_i + N$, where N is a Gaussian noise $\sim \mathcal{N}(0, \sigma_N^2)$, independent of $Z_{[n]}$, and α_i's are nonnegative weights such that $\sum_{i=1}^n \alpha_i = 1$. It can be shown that the true generalization error is also $2\sigma^2/n$ (see the Appendix).

D. Existing Generalization Error Bounds

Xu and Raginsky, motivated by a previous work by Russo and Zou [1], provided a mutual information (MI) based bound on the expected generalization error [2]. Assuming $\ell(w, Z)$ is σ-sub-Gaussian[2] under ξ for all $w \in \mathcal{W}$, then the bound is

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \sqrt{\frac{2\sigma^2}{n} I(W; Z_{[n]})}. \quad (16)$$

One issue with this bound is that it can be vacuous, i.e., the mutual information term can be bounded. Indeed, for the quadratic Gaussian case, it is vacuous when $N = 0$. Bu et al. [9] noticed that the generalization error can be written as

$$\begin{aligned} \mathrm{gen}(\xi, P_{W|Z_{[n]}}) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(\ell(W, \tilde{Z}_i) - \ell(W, Z_i)\right)\right] \quad (17) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(L_\xi(W) - \ell(W, Z_i)\right)\right], \quad (18) \end{aligned}$$

where $\tilde{Z}_i$'s are independent testing data samples that are independent of W. The following bound can then be obtained by bounding each summand

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \leq \frac{1}{n}\sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}, \quad (19)$$

assuming $\ell(\tilde{W}, \tilde{Z})$ is σ-sub-Gaussian, where $\tilde{W}$ and $\tilde{Z}$ are independent but have the same marginal distribution as that in $P_{W,Z_{[n]}}$. This bound improves upon the bound in [2], and it is in general not vacuous. However, for the quadratic Gaussian problem, it leads to an order $\mathcal{O}(1/\sqrt{n})$ bound, which is order-wise loose.

Steinke and Zakynthinou [10] introduced a conditional-mutual-information-based generalization error bound. We will not provide the precise bound here, but it can be shown straightforwardly that their bound leads to an order $\mathcal{O}(1)$ bound, which is order-wise loose. Different improvements on this conditional mutual information bound have been proposed [11], [14], [15], however, in the quadratic Gaussian problem, they led to either $\mathcal{O}(1)$ or $\mathcal{O}(1/\sqrt{n})$ bounds, thus also order-wise loose. Details can be found in [14].

Zhou et al. [16] proposed a chaining technique to tighten the generalization error bound, and showed that with a specially constructed chain in the quadratic Gaussian problem, the bound in [9] can be tightened to the order $\mathcal{O}(1/n)$, but with a loose constant factor. In a more recent work [13],

[2]We call a distribution σ-sub-Gaussian if it has a variance proxy of σ^2.

Wu et al. proposed a new bound assuming the function $r(\tilde{W},\tilde{Z}) = \ell(\tilde{W},\tilde{Z}) - \ell(w^*,\tilde{Z})$ is σ^2-sub-Gaussian, where w^* is the optimal solution of the true risk. For the quadratic Gaussian problem, this bound is asymptotically optimal,[3] but not optimal for finite n. Moreover, the function $r(\tilde{W},\tilde{Z})$ relies on the optimal solution w^*. A more detailed summary of the quadratic Gaussian location problem can be found in [34].

III. A New Information-Theoretic Generalization Error Bound

The new information-theoretic generalization error bound is summarized in the following theorem.

Theorem 1: Let $F_i = L_\xi(W) - \ell(W,Z_i)$, then we have

$$\begin{aligned}\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{Z_i}}\left[\inf_{\lambda>0}\frac{\mathrm{KL}\big(P_{W|Z_i}\|Q^i_W\big)+\Lambda_{F_i|Z_i,Q^i_W}(\lambda)}{\lambda}\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{Z_i}}\Big[\Lambda^{*-1}_{F_i|Z_i,Q^i_W}\big(\mathrm{KL}\big(P_{W|Z_i}\|Q^i_W\big)\big)\Big],\end{aligned} \tag{20}$$

for any $Q^i_{W,Z_i} = Q^i_W P_{Z_i}$, $i = 1,2,\ldots,n$, i.e., a distribution Q^i where W is independent of Z_i.

The reference distribution Q can in fact be optimized, which would provide the tightest bound for a fixed learning algorithm. This bears certain resemblance to those used in [35] which considers the computation of tight generalization bound using the PAC-Bayesian approach.

Proof: We start from (18), and consider each summand on the right-hand side

$$\begin{aligned}&\mathbb{E}_{P_{W,Z_i}}\big[L_\xi(W) - \ell(W,Z_i)\big]\\ &= \mathbb{E}_{P_{Z_i}}\Big[\mathbb{E}_{P_{W|Z_i}}\big((L_\xi(W)-\ell(W,Z_i)\big|Z_i\big)\Big]\\ &\le \mathbb{E}_{P_{Z_i}}\Bigg[\inf_{\lambda>0}\frac{\mathrm{KL}\big(P_{W|Z_i}||Q^i_W\big)+\Lambda_{F_i|Z_i,Q^i_W}(\lambda)}{\lambda}\\ &\qquad + \mathbb{E}_{Q^i_W}\Big(\Big(L_\xi(W)-\ell(W,Z_i)\Big|Z_i\Big)\Bigg]\\ &= \mathbb{E}_{P_{Z_i}}\left[\inf_{\lambda>0}\frac{\mathrm{KL}\big(P_{W|Z_i}||Q^i_W\big)+\Lambda_{F_i|Z_i,Q^i_W}(\lambda)}{\lambda}\right],\end{aligned} \tag{21}$$

where the first equality is by the tower rule, the inequality is by (9), and the second equality is due to (11). Summing over i gives the bound stated in the theorem. ∎

As will be shown in the next section, this bound is exactly tight for the quadratic Gaussian problem, and therefore, it can be viewed as a tight bound in the sense that it cannot be strictly improved in a uniform manner, either in terms of the constant or in the scaling. The proposed bound has operations in the order of infimum, expectation, and summation, where the summation and the expectation are exchangeable without hurting the tightness. We believe that taking the infimum inside is the key to the tight bound, since it allows the bound to fully leverage the flexibility of optimizing λ. To make this explicit, we give the following corollaries.

Corollary 1: Let $F_i = L_\xi(W) - \ell(W,Z_i)$, then we have

$$\begin{aligned}\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) &\le \frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda>0}\mathbb{E}\left[\frac{\mathrm{KL}\big(P_{W|Z_i}\|Q^i_W\big)+\Lambda_{F_i,Q^i_W}(\lambda)}{\lambda}\right]\\ &\le \inf_{\lambda>0}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathrm{KL}\big(P_{W|Z_i}\|Q^i_W\big)+\Lambda_{F_i,Q^i_W}(\lambda)}{\lambda}\right]\right],\end{aligned} \tag{22}$$

for any $Q^i_{W,Z_i} = Q^i_W P_{Z_i}$, $i = 1,2,\ldots,n$.

The first inequality is obtained by exchanging expectation and infimum, and the second is obtained by exchanging summation and infimum.

Corollary 2: Let $F_i = L_\xi(W) - \ell(W,Z_i)$, then we have

$$\begin{aligned}\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) &\le \mathbb{E}\inf_{\lambda>0}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathrm{KL}\big(P_{W|Z_i}\|Q^i_W\big)+\Lambda_{F_i,Q^i_W}(\lambda)}{\lambda}\right]\right]\\ &\le \inf_{\lambda>0}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathrm{KL}\big(P_{W|Z_i}\|Q^i_W\big)+\Lambda_{F_i,Q^i_W}(\lambda)}{\lambda}\right]\right]\end{aligned} \tag{23}$$

for any $Q^i_{W,Z_i} = Q^i_W P_{Z_i}$, $i = 1,2,\ldots,n$.

The first inequality is obtained by exchanging expectation and summation, and the second by exchanging infimum and expectation. The second bounds in Corollaries 1 and 2 are the same, while the first bounds are not directly comparable.

Notice that when $Q^i_{W,Z_i} = P_W \otimes P_{Z_i}$, i.e., the product of the marginals of P_{W,Z_i}, we have $\mathbb{E}[\mathrm{KL}(P_{W|Z_i}\|Q^i_W)] = I(W;Z_i)$. This leads to the following corollary.

Corollary 3: Let $F_i = L_\xi(W) - \ell(W,Z_i)$, then we have

$$\begin{aligned}\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) &\le \frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda>0}\left[\frac{I(W;Z_i)+\mathbb{E}\Lambda_{F_i,P_W}(\lambda)}{\lambda}\right]\\ &\le \frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda>0}\left[\frac{I(W;Z_i)+\Lambda_{F_i,P_WP_{Z_i}}(\lambda)}{\lambda}\right],\\ &= \frac{1}{n}\sum_{i=1}^{n}\Lambda^{*-1}_{F_i,P_WP_{Z_i}}(I(W;Z_i))\end{aligned} \tag{24}$$

where the second inequality is due to the concavity of the $\ln(\cdot)$ function.

By exchanging the infimum and the summation, we straightforwardly obtain further that

$$\begin{aligned}\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) &\le \inf_{\lambda>0}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(W;Z_i)+\mathbb{E}\Lambda_{F_i,P_W}(\lambda)}{\lambda}\right]\right]\\ &\le \inf_{\lambda>0}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(W;Z_i)+\Lambda_{F_i,P_WP_{Z_i}}(\lambda)}{\lambda}\right]\right].\end{aligned} \tag{25}$$

The second bound in (24) is in fact quite similar to the main theorem in [9]. However, there is a major difference even when we assume the reference distribution Q is the same as

[3]The bound is only asymptotically optimal, (in fact, only asymptotically valid) since one of the inequalities is replaced by an approximation that only holds in an asymptotic manner to yield the bound. Strictly speaking, their bound can be stated as follows: for any $\epsilon > 0$, for sufficiently large n, the generalization error $\le 2(1+\epsilon)\sigma^2/n$ in this quadratic Gaussian problem.

the product of the marginals in P: the function F we choose to bound is different. Specifically, the bound proposed by Bu et al. has the exact same form as (24), however with the function F_i being the loss function $\ell(W, Z_i)$, while we use the generalization error as the function, i.e., $F_j = L_\xi(W) - \ell(W, Z_i)$.

When the function F is conditional σ_{Z_i}-sub-Gaussian with respect to the distribution Q_W, we have as a consequence $\Lambda_{F_i, Q^i_W}(\lambda) \le \sigma^2_{Q Z_i}\lambda^2$. The following corollary is then immediate.

Corollary 4: Let $F_i = L_\xi(W) - \ell(W, Z_i)$. If F_i is conditional $\sigma_{Q_{z_i}}$-sub-Gaussian for each $Z_i = z_i$ with respective to Q^i_W then

$$\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\sqrt{\text{KL}\big(P_{W|Z_i}\|Q^i_W\big)\sigma^2_{QZ_i}} \le \frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbb{E}\Big[\text{KL}\big(P_{W|Z_i}\|Q^i_W\big)\sigma^2_{QZ_i}\Big]}. \quad (26)$$

for any Q^i_W such that W is independent of Z_i for $i = 1, 2, \ldots, n$.

IV. The Canonical Quadratic Gaussian Problem Revisited

With the new generalization error bounds derived in the previous section, we are now ready to revisit the canonical quadratic Gaussian (location) problem.

A. Exactly Tight Bounds for the Quadratic Gaussian Problem

The expected generalization error of interest in the quadratic Gaussian problem is

$$\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\sigma^2 + \mu^2 - Z_i^2 + 2(Z_i - \mu)W|Z_i\Big]\right]. \quad (27)$$

For any fixed i, define

$$F_i = f_{Z_i}(W) \triangleq \sigma^2 + \mu^2 - Z_i^2 + 2(Z_i - \mu)W. \quad (28)$$

Note the conditional distribution

$$W|Z_i \overset{P}{\sim} \mathcal{N}\left(\mu + \alpha_i(Z_i - \mu), \sum_{j\neq i}\alpha_j^2\sigma^2 + \sigma_N^2\right). \quad (29)$$

We will choose the reference distribution Q^i_W as

$$W \overset{Q^i_W}{\sim} \mathcal{N}\left(\mu, \sum_{j\neq i}\alpha_j^2\sigma^2 + \sigma_N^2\right), \quad (30)$$

which is indeed independent of Z_i.

Remark. In the reference distribution Q^i_{W,Z_i}, W and Z_i are independent, and the marginal distribution Q^i_W is not the same as that marginalized from $P_{W,Z_{[n]}}$. More specifically, the latter is in fact

$$P_W \sim \mathcal{N}\left(\mu, \sum_{i=1}^{n}\alpha_i^2\sigma^2 + \sigma_N^2\right),$$

which can be compared with (30).

With these conditional distributions, we can derive that (see the Appendix)

$$\begin{aligned}\text{KL}\Big(P_{W|Z_i}\|Q^i_{W|Z_i}\Big) &= \text{KL}\big(P_{W|Z_i}\|Q^i_W\big) \\ &= \alpha_i^2(Z_i - \mu)^2\frac{1}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2}; \\ \Lambda_{F_i, Q^i_{W|Z_i}}(\lambda) &= \Lambda_{F_i, Q^i_W}(\lambda) \\ &= 2\lambda^2(Z_i - \mu)^2\left(\sum_{j\neq i}\alpha_j^2\sigma^2 + \sigma_N^2\right). \end{aligned} \quad (31)$$

Therefore

$$\begin{aligned}\mathbb{E}\big[\text{KL}\big(P_{W|Z_i}\|Q^i_W\big)\big] &= \alpha_i^2\sigma^2\frac{1}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2}; \\ \mathbb{E}\Big[\Lambda_{F_i, Q^i_W}(\lambda)\Big] &= 2\lambda^2\sigma^2\left(\sum_{j\neq i}\alpha_j^2\sigma^2 + \sigma_N^2\right). \end{aligned} \quad (32)$$

Applying the first bound in Corollary 1, we obtain

$$\begin{aligned}&\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) \\ &\le \frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda>0}\mathbb{E}\left[\frac{\text{KL}\big(P_{W|Z_i}\|Q^i_W\big) + \Lambda_{F_i, Q^i_W}(\lambda)}{\lambda}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda>0}\left[\frac{\mathbb{E}\big[\text{KL}\big(P_{W|Z_i}\|Q^i_W\big)\big] + \mathbb{E}[\Lambda_{F_i, Q^i_W}(\lambda)]}{\lambda}\right] \\ &= \frac{2\sigma^2}{n}, \end{aligned} \quad (33)$$

where the last equality is by choosing the minimizer λ_i^* as

$$\lambda_i^* = \frac{\alpha_i}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2}. \quad (34)$$

Therefore, the first bound in Corollary 1 leads to a tight generalization error bound for this setting.

Remark. Recall the equality condition in (12). With the given $P_{W|Z_i}$ and Q^i_W, we have that

$$\ln\frac{dP}{dQ} = \frac{2\alpha_i(Z_i - \mu)W - \alpha_i^2(Z_i - \mu)^2 - 2\mu\alpha_i(Z_i - \mu)}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2}. \quad (35)$$

With (28), it is seen that the condition given in (12) is indeed satisfied, with

$$\lambda = \frac{\alpha_i}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2}, \quad (36)$$

$$b = -\frac{\alpha_i^2(Z_i - \mu)^2 + 2\mu\alpha_i(Z_i - \mu)}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2}. \quad (37)$$

This choice of λ is in fact exactly the optimizing solution in (34). Conversely, the distribution Q we chose can be viewed as obtained through the condition (12) (or equivalently $dQ \propto \exp(-\lambda f)\, dP$), with the parameter λ chosen to maintain the independence between W and Z_i as required in Theorem 1.

In contrast to the tight bound derived from the first bound in Corollary 1, the second bound in Corollary 1 and the first bound in Corollary 2 are not tight for general assignments of α_i's, due to the fact that the optimal λ_i^* is index-dependent. In the extreme case, consider setting $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2, 3, \ldots, n$. Then the second bound in Corollary 1 gives

$$\begin{aligned}\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) &= \frac{1}{n}\inf_{\lambda>0}\left[\frac{\frac{\sigma^2}{2\sigma_N^2} + 2(n-1)\sigma^2\big(\sigma^2+\sigma_N^2\big)\lambda^2 + 2\sigma^2\sigma_N^2\lambda^2}{\lambda}\right] \\ &= \frac{2\sigma^2}{n}\sqrt{\frac{2(n-1)\big(\sigma^2+\sigma_N^2\big)+\sigma_N^2}{2\sigma_N^2}}, \end{aligned} \tag{38}$$

which is of order $\mathcal{O}(1/\sqrt{n})$. However, when $\alpha_i = 1/n$, this dependence disappears and the loosened bounds also become tight. Indeed, consider the second bound in Corollary 1 for this case, we have

$$\begin{aligned}\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) &= \frac{1}{n}\inf_{\lambda>0}\left[\frac{\sum_{i=1}^n\Big(\mathbb{E}\big[\text{KL}\big(P_{W|Z_i}\|Q_W^i\big)\big] + \mathbb{E}\Big[\Lambda_{F_i,Q_W^i}(\lambda)\Big]\Big)}{\lambda}\right] \\ &= \frac{2\sigma^2}{n}, \end{aligned} \tag{39}$$

where the last step is obtained by choosing

$$\lambda^* = \frac{\alpha_i}{2\sum_{j\neq i}\alpha_j^2\sigma^2 + 2\sigma_N^2} = \frac{n}{2(n-1)\sigma^2 + 2n\sigma_N^2}. \tag{40}$$

Remark. The additive noise N in the algorithm $W = \sum_{i=1}^n \alpha_i Z_i + N$ makes it a randomized algorithm, but it does not cause any essential difference in our bound. We included the noise here mostly to enlarge the set of problems that the proposed generalization error bound is tight. In other words, the proposed bound is not only tight for one particular algorithm of $\alpha_i = 1/n$ and $\sigma_N^2 = 0$, but also a class of algorithms with different α_i's and σ_N^2.

B. Looseness of Mutual Information Based Bounds

One remaining question in the quadratic Gaussian problem is whether we can obtain tight or asymptotically tight generalization error bounds using mutual-information-based bounds. To understand this issue, we consider the bounds in Corollary 3 assuming the coefficients $\alpha_i = 1/n$ for $i = 1, 2, \ldots, n$. Note that in this case, the choice of the reference distribution Q_W^i is fixed as the marginal of P_W.

The various terms we need when applying Corollary 3 in this setting can be shown to be (see the Appendix)

$$\begin{aligned} I(W; Z_i) &= \frac{1}{2}\log\frac{n}{n-1} \\ \mathbb{E}\Lambda_{F_i,Q_W^i}(\lambda) &= \frac{2\sigma^4(n-1)}{n^2}\lambda^2 \\ \Lambda_{F_i,Q_{W,Z_i}^i}(\lambda) &= \lambda\sigma^2 - \frac{1}{2}\log\left[1 - 2\left(\frac{2\lambda^2\sigma^4}{n} - \lambda\sigma^2\right)\right]. \end{aligned}$$

With these quantities, it follows that the first bound in Corollary 3 is

$$\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) \le \frac{2\sigma^2}{n}\sqrt{\left(\log\frac{n}{n-1}\right)(n-1)}. \tag{41}$$

The bound is of order $\mathcal{O}(1/n)$; in fact, it is asymptotically optimal in the sense that it approaches $\frac{2\sigma^2}{n}$. Therefore, the first mutual-information-based bound in Corollary 3 does not lose the tightness in a significant manner compared to the KL-based bound of those in Corollaries 1 and 2.

The second bound in Corollary 3 has the form

$$\begin{aligned}\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) &\le \sigma^2 \\ &+ \inf_{\lambda>0}\left[\frac{1}{2\lambda}\log\frac{n}{n-1} - \frac{1}{2\lambda}\log\left[1 - 2\left(\frac{2\lambda^2\sigma^4}{n} - \lambda\sigma^2\right)\right]\right], \end{aligned} \tag{42}$$

for any $\delta \in (0, 1/2]$, and any $\epsilon > 0$, by choosing $\lambda = 1/(2n^\delta\sigma^2)$, it can be seen that for sufficiently large n, we have $\text{gen}(\xi, P_{W|Z_{[n]}}) \le (1+\epsilon)\frac{2\sigma^2}{n^{1-\delta}}$. Therefore, the bound can be also viewed as asymptotically optimal.

Similarly, we can apply the bounds in (25). Since in this case, the optimal choice of λ does not depend on the index-i, they are also asymptotically optimal. It should be noted that when the weight coefficients α_i's are not chosen to be uniform, then the optimal λ becomes dependent on the index i, and the bounds in (25) will be looser, in a similar manner as that for the KL-based bounds.

From the discussion on both the KL-based bound and the mutual-information-based bounds, it appears that the order-wise looseness of the existing bounds mainly stems from the choice of the function to apply the change of measure inequality, i.e., $\ell(W, Z_i)$ or $\ell(W, \tilde{Z}) - \ell(W, Z_i)$. It is seen that the second quality is intuitively more centered, and therefore, the variance proxy is considerably lower than the former, assuming that they are both sub-Gaussian. In the canonical Gaussian setting, this difference is critical to make the information-theoretic bounds tight or asymptotically tight. We expect the same effect will manifest in other problem settings, though without the ground truth and the statistical models, this conjecture is difficult to verify precisely.

V. The Vector Quadratic Gaussian Location Problem

Let us consider the vector version of the quadratic Gaussian location problem. Let the data samples be $Z_1, Z_2, \ldots, Z_n \overset{i.i.d.}{\sim} \xi = \mathcal{N}(\mu, \Sigma)$, i.e., each Z_i is a d-dimensional random Gaussian vector. The sample-average algorithm again chooses the following hypothesis $W = \sum_{i=1}^n \alpha_i Z_i + N$, where α_i's are nonnegative weights such that $\sum_{i=1}^n \alpha_i = 1$, and N is a Gaussian noise vector $\sim \mathcal{N}(0, \sigma_N^2\mathbf{I})$. Instead of considering the standard mean squared error, let us consider a more general quadratic distortion measure $\|x\|_A^2 = x^T Ax$, based on a symmetric positive definite matrix A, for which we have

$$\begin{aligned}\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) &= \mathbb{E}\Bigg[\big(\tilde{Z} - W\big)^T A\big(\tilde{Z} - W\big) \\ &\qquad - \frac{1}{n}\sum_{i=1}^n (Z_i - W)^T A(Z_i - W)\Bigg] \\ &= \frac{1}{n}\sum_{i=1}^n\big[\text{Tr}\big(A\big(\Sigma + \mu\mu^T\big)\big) \\ &\qquad - \mathbb{E}\big(Z_i^T A Z_i - 2(Z_i - \mu)^T AW\big)\big]. \end{aligned} \tag{43}$$

It can be shown that the generalization error of this setting is $\frac{2\mathrm{Tr}(A\Sigma)}{n}$.

One would expect that the result on the scalar setting could be generalized to this setting to obtain tight bounds, however, we shall illustrate the critical condition (12) is in fact rather stringent. To obtain tight results in this setting, one has to apply the bound in a different manner and the tightness is dependent on the decomposition of the loss function.

A. Generalization Error Bounds via Theorem 1

Let us follow the footsteps of the scalar case, and define

$$F_i = \mathrm{Tr}\big(A(\Sigma + \mu\mu^T)\big) - \big(Z_i^T A Z_i - 2(Z_i - \mu)^T A W\big). \quad (44)$$

The conditional distribution is

$$W|Z_i \overset{P}{\sim} \mathcal{N}\left(\mu + \alpha_i(Z_i - \mu), \sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right). \quad (45)$$

We will choose the reference distribution Q_W^i as

$$W \overset{Q_W^i}{\sim} \mathcal{N}\left(\mu, \sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right), \quad (46)$$

which is independent of Z_i.

With these conditional distributions, we can derive (see appendix) that

$$\begin{aligned} &\mathrm{KL}\big(P_{W|Z_i} || Q_W^i\big) \\ &= \frac{\alpha_i^2}{2}\left[(Z_i - \mu)^T \left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right)^{-1} (Z_i - \mu)\right] \quad (47) \\ &\Lambda_{F_i, Q_W^i}(\lambda) \\ &= 2\lambda^2 (Z_i - \mu)^T A \left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right) A (Z_i - \mu). \quad (48) \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{E}\big[\mathrm{KL}\big(P_{W|Z_i} || Q_W^i\big)\big] &= \frac{\alpha_i^2}{2} \mathrm{Tr}\left[\left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right)^{-1} \Sigma\right] \\ \mathbb{E}\Big[\Lambda_{F_i, Q_W^i}(\lambda)\Big] &= 2\lambda^2 \mathrm{Tr}\left[A\left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right) A \Sigma\right]. \quad (49) \end{aligned}$$

At this point, it is clear that the bounds can not be further simplified under general choices of α_i's, A, and σ_N^2. Next, we consider three special cases:

- $\sigma_N^2 = 0$ and $A = \mathbf{I}$: In this case, we have

$$\begin{aligned} \mathbb{E}\big[\mathrm{KL}\big(P_{W|Z_i} || Q_W^i\big)\big] &= \frac{d\alpha_i^2}{2\sum_{j\neq i} \alpha_j^2}; \\ \mathbb{E}\Big[\Lambda_{F_i, Q_W^i}(\lambda)\Big] &= 2\lambda^2 \left(\sum_{j\neq i} \alpha_j^2\right) \mathrm{Tr}\big[\Sigma^2\big]. \quad (50) \end{aligned}$$

Applying the first bound in Corollary 1, we obtain

$$\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) \leq \frac{2}{n}\sqrt{d\,\mathrm{Tr}\big[\Sigma^2\big]}. \quad (51)$$

As a reference, the true generalization error in this setting is in fact $\frac{2}{n}\mathrm{Tr}[\Sigma]$, i.e., the bound is loose using this bounding approach.

- $\sigma_N^2 = 0$ and $A = \Sigma^{-1}$: In this case, we have

$$\begin{aligned} \mathbb{E}\big[\mathrm{KL}\big(P_{W|Z_i} || Q_W^i\big)\big] &= \frac{d\alpha_i^2}{2\sum_{j\neq i} \alpha_j^2}; \\ \mathbb{E}\Big[\Lambda_{F_i, Q_W^i}(\lambda)\Big] &= 2d\lambda^2 \left(\sum_{j\neq i} \alpha_j^2\right). \quad (52) \end{aligned}$$

Applying the first bound in Corollary 1, we obtain

$$\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) \leq \frac{2d}{n}. \quad (53)$$

For this case, the true generalization error is indeed fact $\frac{2d}{n}$, i.e., the bound is tight using this bounding approach. This setting is however a trivial setting, where the loss function essentially decomposes the vector into i.i.d. components.

- $A = \mathbf{I}$, and $\Sigma = \sigma^2 \mathbf{I}$: In this case, we have

$$\begin{aligned} \mathbb{E}\big[\mathrm{KL}\big(P_{W|Z_i} || Q_W^i\big)\big] &= \frac{d\alpha_i^2 \sigma^2}{2\sum_{j\neq i} \alpha_j^2 \sigma^2 + 2\sigma_N^2}; \\ \mathbb{E}\Big[\Lambda_{F_i, Q_W^i}(\lambda)\Big] &= 2\lambda^2 \left(\sum_{j\neq i} \alpha_j^2 \sigma^2 + \sigma_N^2\right) d\sigma^2. \quad (54) \end{aligned}$$

Applying the first bound in Corollary 1, we obtain

$$\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big) \leq \frac{2}{n} d\sigma^2. \quad (55)$$

The true generalization error in this setting is indeed the same, i.e., the bound is also tight for this special case.

It is seen that in general the bounds derived from the proposed bounds given in Theorem 1 are not tight, but can yield tight bounds for certain special cases.

Remark. Recall the equality condition in (12). With the given $P_{W|Z_i}$ and Q_W^i, we have that

$$\begin{aligned} \ln \frac{dP}{dQ} &= -(W - \mu - \alpha_i(Z_i - \mu))^T \left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right)^{-1} \\ &\quad \cdot (W - \mu - \alpha_i(Z_i - \mu)) \\ &\quad + (W - \mu)^T \left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right)^{-1} (W - \mu) \\ &= 2\alpha_i (Z_i - \mu)^T \left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right)^{-1} (W - \mu). \end{aligned} \quad (56)$$

With (28), it is seen that the condition given in (12) can be satisfied when

$$\left(\sum_{j\neq i} \alpha_j^2 \Sigma + \sigma_N^2 \mathbf{I}\right)^{-1} \propto A, \quad (57)$$

which indeed holds for the latter two cases discussed above. However, this relation does not hold under general Σ, σ_N^2, and A choices, and bounds derived from Theorem 1 will in general be loose.

B. Generalization Error Bounds via Loss Function Decomposition

Recall the loss function in general has the form $\ell : \mathcal{W}\times\mathcal{Z} \to \mathbb{R}$. We say the functions $\ell_j : \mathcal{W}\times\mathcal{Z} \to \mathbb{R}$, $j = 1, 2, \dots, d$, and the functions $\phi_j : \mathcal{W} \to \mathcal{W}_j$, $j = 1, 2, \dots, d$ form a decomposition of the loss function ℓ, if

$$\ell(w, z) = \sum_{j=1}^{d} \ell_j\big(\phi_j(w), z\big), \tag{58}$$

for any $(w, z) \in \mathcal{W}\times\mathcal{Z}$. Clearly, the loss function we have adopted for the vector Gaussian location problem has a decomposition with

$$\begin{aligned}\ell_j(w, Z_i) &= \lambda_j (W - Z_i)^T U_j U_j^T (W - Z_i)\\ &= \Big(\sqrt{\lambda_j}U_j^T W - \sqrt{\lambda_j}U_j^T Z_i\Big)^T \Big(\sqrt{\lambda_j}U_j^T W - \sqrt{\lambda_j}U_j^T Z_i\Big),\end{aligned} \tag{59}$$

where UDU^T is the eigenvalue decomposition of A, λ_j is the j-th diagonal item of D, U_j is the j-th column of U, and $\phi_j = \sqrt{\lambda_j}U_j^T W$.

For a decomposition of the loss function, we have the following generalization of Theorem 1.

Theorem 2: Let $F_{i,j} = L_{j,\xi}(\phi_j(W)) - \ell_j(\phi_j(W), Z_i)$, then we have

$$\begin{aligned}&\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big)\\ &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbb{E}_{P_{Z_i}}\left[\inf_{\lambda>0}\left[\frac{\mathrm{KL}\Big(P_{\phi_j(W)|Z_i}\|Q^i_{\phi_j(W)}\Big)}{\lambda} + \frac{\Lambda_{F_{i,j}|Z_i,Q^i_W}(\lambda)}{\lambda}\right]\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbb{E}_{P_{Z_i}}\left[\Lambda^{*-1}_{F_{i,j}|Z_i,Q^i_W}\Big(\mathrm{KL}\Big(P_{\phi_j(W)|Z_i}\|Q^i_{\phi_j(W)}\Big)\Big)\right],\end{aligned} \tag{60}$$

for $Q^i_{\phi_j(W),Z_i} = Q^i_{\phi_j(W)}P_{Z_i}$, $i = 1, 2, \dots, n$, that is induced by any $Q^i_{W,Z_i} = Q^i_W P_{Z_i}$, $i = 1, 2, \dots, n$, i.e., a distribution Q^i where W is independent of Z_i.

We omit its proof since it is almost identical to that of Theorem 1. It should be noted that the variational representation inequality is applied on the marginalized distribution $P_{\phi_j(W)|Z_i}$ and $Q^i_{\phi_j(W)}$, however since $Q^i_{\phi_j(W)}$ is induced by Q^i_W, we have $\Lambda_{F_{i,j}|Z_i,Q^i_W}(\lambda) = \Lambda_{F_{i,j}|Z_i,Q^i_{\phi(W)}}(\lambda)$. We provide the following corollary in order to tackle the vector Gaussian setting. A corollary similar to Corollary 2 can also be written, but it is omitted here for conciseness.

Corollary 5: Let $F_{i,j} = L_{j,\xi}(\phi_j(W)) - \ell_j(\phi_j(W), Z_i)$, then we have

$$\begin{aligned}&\mathrm{gen}\big(\xi, P_{W|Z_{[n]}}\big)\\ &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\inf_{\lambda>0}\mathbb{E}\left[\frac{\mathrm{KL}\Big(P_{\phi_j(W)|Z_i}\|Q^i_{\phi_j(W)}\Big)}{\lambda} + \frac{\Lambda_{F_{i,j},Q^i_W}(\lambda)}{\lambda}\right]\\ &\le \inf_{\lambda>0}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbb{E}\left[\frac{\mathrm{KL}\Big(P_{\phi_j(W)|Z_i}\|Q^i_{\phi_j(W)}\Big)}{\lambda} + \frac{\Lambda_{F_{i,j},Q^i_W}(\lambda)}{\lambda}\right]\right],\end{aligned} \tag{61}$$

for $Q^i_{\phi_j(W),Z_i} = Q^i_{\phi_j(W)}P_{Z_i}$, $i = 1, 2, \dots, n$, that is induced by any $Q^i_{W,Z_i} = Q^i_W P_{Z_i}$, $i = 1, 2, \dots, n$.

Equipped with the new bounds above, let us revisit the vector setting. This time, let us define

$$\begin{aligned}F_{i,j} &= \lambda_j \mathrm{Tr}\Big(U_j U_j^T\big(\Sigma + \mu\mu^T\big)\Big)\\ &\quad - \lambda_j\Big(Z_i^T U_j U_j^T Z_i - 2(Z_i - \mu)^T U_j U_j^T W\Big).\end{aligned} \tag{62}$$

The conditional distribution $P_{\phi_j(W)|Z_i}$ is given as

$$\begin{aligned}P_{\phi_j(W)|Z_i} &= P_{\sqrt{\lambda_j}U_j^\top W|Z_i}\\ &= \mathcal{N}\left(\sqrt{\lambda_j}U_j^\top\mu + \alpha_i\sqrt{\lambda_j}U_j^\top(Z_i - \mu),\ \lambda_j\sum_{j\neq i}\alpha_j^2 U_j^\top\Sigma U_j + \lambda_j\sigma_N^2\right).\end{aligned} \tag{63}$$

We will choose the reference distribution $Q^i_{\phi_j(W)} = Q^i_{\sqrt{\lambda_j}U_j^T W}$ as

$$\sqrt{\lambda_j}U_j^\top W \sim \mathcal{N}\left(\sqrt{\lambda_j}U_j^\top\mu,\ \lambda_j\sum_{j\neq i}\alpha_j^2 U_j^\top\Sigma U_j + \lambda_j\sigma_N^2\right). \tag{64}$$

The divergence term $\mathrm{KL}(P_{\phi_j(W)|Z_i}||Q^i_{\phi_j(W)})$ is therefore

$$\begin{aligned}&\mathrm{KL}\Big(P_{\sqrt{\lambda_j}U_j^\top W|Z_i}||Q^i_{\sqrt{\lambda_j}U_j^\top W}\Big)\\ &= \frac{\alpha_i^2\mathrm{Tr}\Big[U_j U_j^T(Z_i - \mu)(Z_i - \mu)^T\Big]}{2\Big(\sum_{j\neq i}\alpha_j^2 U_j^\top\Sigma U_j + \sigma_N^2\Big)}.\end{aligned} \tag{65}$$

By substituting $A = U_j U_J^T$ in (48), we can obtain that

$$\begin{aligned}&\Lambda_{F_{i,j},Q^i_W}(\lambda)\\ &= 2\lambda_j^2\lambda^2(Z_i - \mu)^T U_j U_j^T\left(\sum_{j\neq i}\alpha_j^2\Sigma + \sigma_N^2\mathbf{I}\right)\cdot U_j U_j^T(Z_i - \mu)\\ &= 2\lambda_j^2\lambda^2\left(\sum_{j\neq i}\alpha_j^2 U_j^T\Sigma U_j + \sigma_N^2\right)\\ &\quad\cdot\mathrm{Tr}\Big[U_j U_j^T(Z_i - \mu)(Z_i - \mu)^T\Big]\end{aligned} \tag{66}$$

Therefore

$$\begin{aligned}\mathbb{E}\Big[\mathrm{KL}\Big(P_{\phi_j(W)|Z_i}||Q^i_{\phi(W)}\Big)\Big] &= \frac{\alpha_i^2\mathrm{Tr}\Big[U_j U_j^T\Sigma\Big]}{2\Big(\sum_{j\neq i}\alpha_j^2 U_j^\top\Sigma U_j + \sigma_N^2\Big)}\\ \mathbb{E}\Big[\Lambda_{F_{i,j},Q^i_W}(\lambda)\Big]&\\ = 2\lambda_j^2\lambda^2\left(\sum_{j\neq i}\alpha_j^2 U_j^T\Sigma U_j + \sigma_N^2\right)&\mathrm{Tr}\Big[U_j U_j^T\Sigma\Big].\end{aligned} \tag{67}$$

Applying the first bound in Corollary 5, we obtain

$$\begin{aligned}
&\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\inf_{\lambda>0}\mathbb{E}\left[\frac{\text{KL}\Big(P_{\phi_j(W)|Z_i}||Q^i_{\phi(W)}\Big)}{\lambda}\right. \\
&\qquad\qquad\qquad\qquad\left.+\frac{\Lambda_{F_{i,j},Q^i_W}(\lambda)}{\lambda}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}2\alpha_j\lambda_j\text{Tr}[U_jU_j^T\Sigma] \\
&= \frac{2}{n}\sum_{j=1}^{d}\lambda_j\text{Tr}\Big[U_jU_j^T\Sigma\Big] \\
&= \frac{2}{n}\text{Tr}\left[\sum_{j=1}^{d}\lambda_jU_jU_j^T\Sigma\right] = \frac{2}{n}\text{Tr}[A\Sigma], \qquad (68)
\end{aligned}$$

which is indeed the true generalization error.

Remark: The decompositions of loss functions are not unique, and clearly certain decomposition can lead to better generalization bounds. A naive decomposition is $\ell_i(w,z) = \frac{1}{d}\ell(w,z)$ and $\phi_j(w) = w$, which does not provide any gain for the vector Gaussian location problem, yet the proposed decomposition can indeed be utilized to yield a tight information-theoretic generalization bound, as shown above. In a sense, decomposition allows us to utilize the probability distribution of a random variable after further processing, and by the data-processing inequality of KL divergence [36], such processing will reduce the KL divergence and potentially yield tighter bounds.

VI. Conclusion

We studied the information-theoretic generalization error bounds, and in particular, focused on the quadratic Gaussian problem. The proposed new bound is shown to be exactly tight for this setting. The most important change from the previous work appears to be the function that we choose to bound, however, the additional introduction of a reference distribution, and the conditional application of the change of measure inequality also contribute to the tightness of the bound. A generalized vector version of the problem is further studied, which inspired a new and refined generalization error bound that relies on the decomposition of the loss functions.

Though we have focused on the quadratic Gaussian setting exclusively in this work, the technique can be applied to the study of noisy and iterative algorithms such as stochastic gradient Langevin dynamics (SGLD), as previously studied in [4], [9], [11], [15]. The key difference from the previous result is that due to the application of the change of measure inequality, our bound relies on the cumulant generating function of a different quantity, or a different sub-Gaussian variance proxy, that likely has a lower value, and therefore the resultant bound is also potentially tighter in that setting. However, due to the more complex statistical dependence induced by the algorithm, it is not clear whether this can drive order-wise gains, and we leave this to a future study.

Gaussian models have had many successes in machine learning research, particularly in the context of Gaussian process [37] and the recent development of Gaussian diffusion models [6], [38], [39], [40], [41]. Therefore, we believe studying the Gaussian settings in the context of machine learning is indeed well-motivated, and will lead to important engineering insights in the future.

Appendix

We can write as follows to derive the exact generalization error for the canonical quadratic Gaussian problem without utilizing the information-theoretical bounds as follows:

$$\begin{aligned}
\text{gen}\big(\xi, P_{W|Z_{[n]}}\big) &= \mathbb{E}\left[\big(\tilde{Z}-W\big)^2 - \frac{1}{n}\sum_{i=1}^{n}(Z_i-W)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\sigma^2+\mu^2-Z_i^2+2(Z_i-\mu)W\Big) \\
&= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_i-\mu)\left(\sum_{j=1}^{n}\alpha_jZ_j+N\right)\right] \\
&= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_i-\mu)\left(\sum_{j=1}^{n}\alpha_j\big(Z_j-\mu\big)+N\right)\right] \\
&= \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_j\mathbb{E}\big[(Z_i-\mu)\big(Z_j-\mu\big)\big] \\
&= \frac{2}{n}\sum_{i=1}^{n}\alpha_i\mathbb{E}\Big[(Z_i-\mu)^2\Big] = \frac{2\sigma^2}{n}. \qquad (69)
\end{aligned}$$

This gives the exact generalization error for this setting.

First, notice that

$$\begin{aligned}
\mathbb{E}_{Q^i_{W|Z_i}}[F_i] &= \mathbb{E}_{Q^i_{W|Z_i}}\Big[\Big(\sigma^2+\mu^2-Z_i^2\Big)+2(Z_i-\mu)W|Z_i\Big] \\
&= \Big(\sigma^2+\mu^2-Z_i^2\Big)+2\mu(Z_i-\mu), \qquad (70)
\end{aligned}$$

since under $Q^iW|Z_i$, W and Z_i are independent, and W has mean μ. Then we can write

$$\begin{aligned}
&\mathbb{E}_{Q^i_{W|Z_i}}\exp\Big[\lambda\Big(\sigma^2+\mu^2-Z_i^2\Big)+2\lambda(Z_i-\mu)W|Z_i\Big] \\
&= \exp\Big[\lambda\Big(\sigma^2+\mu^2-Z_i^2\Big)\Big]\mathbb{E}_{Q^i_{W|Z_i}}\big[\exp(2\lambda(Z_i-\mu)W)|Z_i\big] \\
&= \exp\Big[\lambda\Big(\sigma^2+\mu^2-Z_i^2\Big)\Big]\exp[2\lambda\mu(Z_i-\mu)] \\
&\quad\cdot\exp\left[2\lambda^2(Z_i-\mu)^2\left(\sum_{j\neq i}\alpha_j^2\sigma^2+\sigma_N^2\right)\right], \qquad (71)
\end{aligned}$$

where the second equality is by the moment generating function of Gaussian random variable W distributed according to Q^i_W. It follows then

$$\begin{aligned}
\Lambda_{F_i,Q^i_{W|Z_i}}(\lambda) &= \ln\mathbb{E}_{Q^i_{W|Z_i}}\big[\exp(\lambda F_i)\big]-\lambda\mathbb{E}[F_i] \\
&= 2\lambda^2(Z_i-\mu)^2\left(\sum_{j\neq i}\alpha_j^2\sigma^2+\sigma_N^2\right). \qquad (72)
\end{aligned}$$

First, notice that $\mathbb{E}_{P_W P_{Z_i}}[F_i] = 0$. Then

$$\begin{aligned}
&\mathbb{E}_{P_W P_{Z_i}} \exp\Big(\lambda\Big(\sigma^2 + \mu^2 - Z_i^2\Big) + 2\lambda(Z_i - \mu)W\Big) \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\exp\Big(\lambda\Big(\sigma^2 + \mu^2 - Z_i^2\Big) + 2\lambda(Z_i - \mu)W\Big)|Z_i\Big]\Big] \\
&= \mathbb{E}\Big[\exp\Big(\lambda(\sigma^2 + \mu^2 - Z_i^2)\Big) \cdot \exp\Big(2\lambda(Z_i - \mu)\mu \\
&\qquad + 2\lambda^2(Z_i - \mu)^2\sigma^2/n\Big)|Z_i\Big], \qquad (73)
\end{aligned}$$

where the first equality is by the tower rule, and the second step is by using the moment generating function of the Gaussian random variable W. Rearranging the terms gives

$$\begin{aligned}
&\mathbb{E}_{P_W P_{Z_i}} \exp\Big(\lambda\Big(\sigma^2 + \mu^2 - Z_i^2\Big) + 2\lambda(Z_i - \mu)W\Big) \\
&= \exp\Big(\lambda\sigma^2\Big)\mathbb{E}\exp\Big[\Big(\frac{2\lambda^2\sigma^2}{n} - \lambda\Big)(Z_i - \mu)^2\Big] \\
&= \exp\Big(\lambda\sigma^2\Big)\Big(1 - 2\Big(\frac{2\lambda^2\sigma^2}{n} - \lambda\Big)\sigma^2\Big)^{-1/2}, \qquad (74)
\end{aligned}$$

where the last equality is by the moment generating function of the χ^2 random variable of degree one. Taking the logarithm on the right-hand side gives the expression for $\Lambda_{F_i, P_W P_{Z_i}}(\lambda)$.

Similar to the scalar case, notice that

$$\begin{aligned}
&\mathbb{E}_{Q^i_{W|Z_i}}[F_i] \\
&= \mathbb{E}_{Q^i_{W|Z_i}}\big[\mathrm{Tr}\big(A\big(\Sigma + \mu\mu^T\big)\big) \\
&\qquad - \big(Z_i^T A Z_i - 2(Z_i - \mu)^T A W\big)|Z_i\big] \\
&= \big(\mathrm{Tr}\big(A\big(\Sigma + \mu\mu^T\big)\big) - Z_i^T A Z_i\big) + 2(Z_i - \mu)^T A\mu. \quad (75)
\end{aligned}$$

We can then write the exponential term in $\Lambda_{F_i, Q^i_{W|Z_i}}(\lambda)$

$$\begin{aligned}
&\mathbb{E}_{Q^i_{W|Z_i}} \exp\big[\lambda \mathrm{Tr}\big(A\big(\Sigma + \mu\mu^T\big)\big) \\
&\qquad - \lambda\big(Z_i^T A Z_i - 2(Z_i - \mu)^T A W\big)|Z_i\big] \\
&= \exp\big[\lambda \mathrm{Tr}\big(A\big(\Sigma + \mu\mu^T\big)\big) - \lambda Z_i^T A Z_i\big] \\
&\qquad \cdot \mathbb{E}_{Q^i_{W|Z_i}}\big[\exp\big(2\lambda(Z_i - \mu)^T A W\big)|Z_i\big] \\
&= \exp\big[\lambda \mathrm{Tr}\big(A\big(\Sigma + \mu\mu^T\big)\big) - \lambda Z_i^T A Z_i\big] \\
&\quad \cdot \exp\big[2\lambda(Z_i - \mu)^T A\mu\big] \\
&\quad \cdot \exp\left[2\lambda^2(Z_i - \mu)^T A\left(\sum_{j\neq i}\Big(\alpha_j^2\Sigma\Big) + \sigma_N^2\mathbf{I}\right)A^T(Z_i - \mu)\right] \qquad (76)
\end{aligned}$$

where the second equality follows standard manipulation of Gaussian integration. It follows then

$$\begin{aligned}
&\Lambda_{F_i, Q^i_{W|Z_i}}(\lambda) = \ln \mathbb{E}Q^i_{W|Z_i}\big[\exp(\lambda F_i)\big] - \lambda\mathbb{E}[F_i] \\
&= 2\lambda^2(Z_i - \mu)^T A\left(\sum_{j\neq i}\Big(\alpha_j^2\Sigma\Big) + \sigma_N^2\mathbf{I}\right)A^T(Z_i - \mu). \quad (77)
\end{aligned}$$

References

[1] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proc. Artif. Intell. Statist.*, 2016, pp. 1232–1240.

[2] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2524–2533.

[3] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7234–7243.

[4] A. Pensia, V. Jog, and P.-L. Loh, "Generalization error bounds for noisy, iterative algorithms," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 546–550.

[5] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 582–586.

[6] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for SGLD via data-dependent estimates," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11015–11025.

[7] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," *Entropy*, vol. 23, no. 1, p. 126, 2021.

[8] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Information-theoretic analysis for transfer learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2819–2824.

[9] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE J. Sel. Areas Commun.*, vol. 1, no. 1, pp. 121–130, May 2020.

[10] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Proc. Conf. Learn. Theory*, 2020, pp. 3437–3452.

[11] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9925–9935.

[12] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *IEEE J. Sel. Areas Commun.*, vol. 1, no. 3, pp. 824–839, Nov. 2020.

[13] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Fast rate generalization error bounds: Variations on a theme," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2022, pp. 43–48.

[14] R. Zhou, C. Tian, and T. Liu, "Individually conditional individual mutual information bound on generalization error," *IEEE Trans. Inf. Theory*, vol. 68, no. 5, pp. 3304–3316, May 2022.

[15] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 1–5.

[16] R. Zhou, C. Tian, and T. Liu, "Stochastic chaining and strengthened information-theoretic generalization bounds," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2022, pp. 690–695.

[17] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8106–8118.

[18] L. P. Barnes, A. Dytso, and H. V. Poor, "Improved information-theoretic generalization bounds for distributed, federated, and iterative learning," *Entropy*, vol. 24, no. 9, p. 1178, 2022.

[19] G. Aminian, Y. Bu, G. W. Wornell, and M. R. Rodrigues, "Tighter expected generalization error bounds via convexity of information measures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2022, pp. 2481–2486.

[20] M. Haghifam, S. Moran, D. M. Roy, and G. K. Dziugiate, "Understanding generalization via leave-one-out conditional mutual information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2022, pp. 2487–2492.

[21] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, "Conditioning and processing: Techniques to improve information-theoretic generalization bounds," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–11.

[22] F. Hellström and G. Durisi, "Evaluated CMI bounds for meta learning: Tightness and expressiveness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 20648–20660.

[23] M. Haghifam, B. Rodríguez-Gálvez, R. Thobaben, M. Skoglund, D. M. Roy, and G. K. Dziugaite, "Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2023, pp. 663–706.

[24] H. Wang, R. Gao, and F. P. Calmon, "Generalization bounds for noisy iterative algorithms using properties of additive noise channels," *J. Mach. Learn. Res.*, vol. 24, pp. 1–43, Jan. 2023.

[25] Z. Wang and Y. Mao, "Tighter information-theoretic generalization bounds from supersamples," 2023, *arXiv:2302.02432*.

[26] Z. Wang and Y. Mao, "On the generalization of models trained with SGD: Information-theoretic bounds and implications," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–27.

[27] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. New York, NY, USA: Wiley, 1991.

[29] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul.–Oct. 1948.

[30] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[31] T. Berger, "Rate-distortion theory," *Wiley Encyclopedia of Telecommunications*. New York, NY, USA: Wiley, 2003.

[32] R. M. Gray, *Source Coding Theory* (Series in Engineering and Computer Science). Norwell, MA, USA: Springer Int., 1989.

[33] A. Picard-Weibel and B. Guedj, "On change of measure inequalities for f-divergences," 2022, *arXiv:2202.05568*.

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

[35] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," in *Proc. Conf. Uncertainty Artif. Intell.*, 2017, pp. 1–10.

[36] Y. Wu, "Information-theoretic methods for high-dimensional statistics," Lecture Notes, Dept. Stat. Data Sci., Yale Univ., New Haven, CT, USA, 2020. [Online]. Available: http://www.stat.yale.edu/ yw562/teaching/it-stats.pdf

[37] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[38] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[39] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[40] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–36.

[41] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8780–8794.