Information-Theoretic Bounds on the Moments of the Generalization Error of Learning Algorithms

Gholamali Aminian[§], Laura Toni, Miguel R. D. Rodrigues

Department of Electronic and Electrical Engineering
University College London
{g.aminian, l.toni, m.rodrigues}@ucl.ac.uk

Abstract—**Generalization error bounds are critical to understanding the performance of machine learning models. In this work, building upon a new bound of the expected value of an arbitrary function of the population and empirical risk of a learning algorithm, we offer a more refined analysis of the generalization behaviour of a machine learning models based on a characterization of (bounds) to their generalization error moments. We discuss how the proposed bounds – which also encompass new bounds to the expected generalization error – relate to existing bounds in the literature. We also discuss how the proposed generalization error moment bounds can be used to construct new generalization error high-probability bounds.**

Index Terms—**Population Risk, Empirical Risk, Generalization Error, Generalization Error Moments, Information Measures**

I. Introduction

Machine learning-based approaches are increasingly adopted to solve various prediction problems in a wide range of applications such as computer vision, speech recognition, speech translation, and many more [1], [2]. In particular, supervised machine learning approaches learn a predictor – also known as a hypothesis – mapping input variables to output variables using some algorithm that leverages a series of input-output examples drawn from some underlying (and unknown) distribution [1]. It is therefore critical to understand the generalization ability of such a predictor, i.e., how the predictor performance on the training set differs from its performance on a testing set (or on the population).

A recent research direction within the information-theoretic and related communities has concentrated on the development of approaches to characterize the generalization error of *randomized learning algorithms*, i.e. learning algorithms map the set of training examples to the hypothesis according to some probability law [3], [4].

The characterization of the generalization ability of randomized learning algorithms has come in two broad flavours. One involves determining a bound to the generalization error that holds on average. For example, building upon pioneering work by Russo and Zou [5], Xu and Raginsky [3] have derived average generalization error bounds involving the mutual information between the training set and the hypothesis. Bu *et al.* [6] have derived tighter average generalization error bounds involving the mutual information between each sample in the training set and the hypothesis. Bounds using chaining mutual information have been proposed in [7]. The upper bounds based on conditional mutual information are proposed in [8] and [9]. Other authors have also constructed information-theoretic based average generalization error bounds using quantities such as α-Réyni divergence, f-divergence, Jensen-Shannon divergences, Wasserstein distances, or maximal leakage (see [10], [11], [12], [13], [14] or [15]).

The other flavour – known as *probably approximately correct (PAC)-Bayesian* bounds and *single-draw* upper bounds – involves determining a bound to the generalization error that holds with high probability. The original PAC-Bayesian generalization error bounds have been characterized via a Kullback-Leibler (KL) divergence (a.k.a. relative entropy) between a prior data-free distribution and a posterior data-dependent distribution on the hypothesis space [16]. Other slightly different PAC-Bayesian generalization error bounds have also been offered in [17], [18], [19], [20] and [21]. A general PAC-Bayesian framework offering high probability bounds on a convex function of the population risk and empirical risk with respect to a posterior distribution has also been provided in [22]. A PAC-Bayesian upper bound by considering a Gibbs data-dependent prior is provided in [23]. Some single-draw upper bounds have been proposed in [3], [10], and [21].

In this paper, we aspire to offer a more refined analysis of the generalization ability of randomized learning algorithms in view of the fact that the generalization error can be seen as a random variable with distribution that depends on randomized algorithm distribution and the data distribution. The analysis of moments of certain quantities arising in statistical learning problems has already been considered in certain works. For example, Russo and and Zou [5] have analysed bounds to certain moments of the error arising in data exploration problems, whereas Dhurandhar and Dobra [24] have analysed bounds to moments of the error arising in model selection problems. Sharper high probability bounds for sums of functions of independent random variables based on their moments, within the context of stable learning algorithms, have also been derived in [25]. However, to the best of our knowledge, a characterization of bounds to the moments of the generalization error of randomized learning algorithms,

[§] The first author is supported by the Royal Society Newton International Fellowship, grant no. NIF\R1 \192656 .

allowing us to capture better how the population risk may deviate from the empirical risk, does not appear to have been considered in the literature.

Our contributions are as follows:

1) First, we offer a general upper bound on the expected value of a function of the population risk and the empirical risk of a randomized learning algorithm expressed via certain information measures between the training set and the hypothesis.
2) Second, we offer upper bounds on the moments of the generalization error of a randomized learning algorithm deriving from the aforementioned general bound in terms of power information and Chi-square information measures. We also propose another upper bound on the second moment of generalization error in terms of mutual information.
3) Third, we show how to leverage the generalization error moment bounds to construct high-probability bounds showcasing how the population risk deviates from the empirical risk associated with a randomized learning algorithm.
4) Finally, we show how the proposed results bound the true moments of the generalization error via a simple numerical example.

We adopt the following notation in the sequel. Upper-case letters denote random variables (e.g., Z), lower-case letters denote random variable realizations (e.g. z), and calligraphic letters denote sets (e.g. $\mathcal{Z}$). The distribution of the random variable Z is denoted by P_Z and the joint distribution of two random variables (Z_1, Z_2) is denoted by $P_{Z_1 Z_2}$. We let $\log(\cdot)$ represent the natural logarithm. We also let $\mathbb{Z}^+$ represent the set of positive integers.

II. Problem Formulation

We consider a standard supervised learning setting where we wish to learn a hypothesis given a set of input-output examples; we also then wish to use this hypothesis to predict new outputs given new inputs.

We model the input data (also known as features) using a random variable $X \in \mathcal{X}$ where $\mathcal{X}$ represents the input set; we model the output data (also known as labels) using a random variable $Y \in \mathcal{Y}$ where $\mathcal{Y}$ represents the output set; we also model input-output data pairs using a random variable $Z = (X, Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where Z is drawn from $\mathcal{Z}$ per some unknown distribution μ. We also let $\mathcal{S} = \{Z_i\}_{i=1}^n$ be a training set consisting of a number of input-output data points drawn i.i.d. from $\mathcal{Z}$ according to μ.

We represent hypotheses using a random variable $W \in \mathcal{W}$ where $\mathcal{W}$ is a hypothesis class. We also represent a randomized learning algorithm via a Markov kernel that maps a given training set $\mathcal{S}$ onto a hypothesis W of the hypothesis class $\mathcal{W}$ according to the probability law $P_{W|S}$.

Let us also introduce a (non-negative) loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ that measures how well a hypothesis predicts an output given an input. We can now define the population risk and the empirical risk given by:

$$L_P(w, \mu) \triangleq \int_{\mathcal{Z}} \ell(w, z)\mu(z)dz \tag{1}$$

$$L_E(w, s) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i) \tag{2}$$

which quantify the performance of a hypothesis w delivered by the randomized learning algorithm on a testing set (population) and the training set, respectively. We can also define the generalization error as follows:

$$\text{gen}(P_{W|S}, \mu) \triangleq L_P(w, \mu) - L_E(w, s) \tag{3}$$

which quantifies how much the population risk deviates from the empirical risk. This generalization error is a random variable whose distribution depends on the randomized learning algorithm probabilistic law along with the (unknown) underlying data distribution. Therefore, an exact characterization of the behaviour of the generalization error – such as its distribution – is not possible for all learning algorithms.

In order to bypass this challenge, our goal in the sequel will be to derive upper bounds to the moments of the generalization error given by:

$$\overline{\text{gen}^m}(P_{W|S}, \mu) \triangleq \mathbb{E}_{P_{W,S}}[(\text{gen}(P_{W|S}, \mu))^m] \tag{4}$$

in terms of various divergences and information-theoretic measures. In particular, we will use the following divergence measures between two distributions P_X and $P_{X'}$ on a common measurable space $\mathcal{X}$:

- The KL divergence given by:
$$D_{KL}(P_X||P_{X'}) \triangleq \int_{\mathcal{X}} P_X(x) \log\left(\frac{P_X(x)}{P_{X'}(x)}\right) dx$$
- The α-Réyni divergence for $1 \leq \alpha$ given by [26]:
$$D_\alpha(P_X||P_{X'}) \triangleq \frac{1}{\alpha - 1} \log\left(\int_{\mathcal{X}} \left(\frac{P_X(x)}{P_{X'}(x)}\right)^\alpha P_{X'}(x)dx\right)$$
- The power divergence of order t given by [27]:
$$D_P^{(t)}(P_X||P_{X'}) \triangleq \int_{\mathcal{X}} \left(\left(\frac{P_X(x)}{P_{X'}(x)}\right)^t - 1\right) P_{X'}(x)dx$$
where $D_\alpha(P_X||P_{X'}) = \frac{\log(D_P^{(t)}(P_X||P_{X'})+1)}{t-1}$ for $\alpha = t$ and $1 \leq \alpha$.
- The Chi-square divergence given by [27]:
$$\chi^2(P_X||P_{X'}) \triangleq \int_{\mathcal{X}} \frac{(P_X(x) - P_{X'}(x))^2}{P_{X'}(x)} dx$$
where $\chi^2(P_X||P_{X'}) = D_P^{(2)}(P_X||P_{X'})$.

We also use the following information measures between two random variables X and X' with joint distribution $P_{XX'}$ and marginals P_X and $P_{X'}$:

- The mutual information given by:
$$I(X; X') \triangleq D_{KL}(P_{X,X'}||P_X \otimes P_{X'})$$

- The power information of order t given by:
$$I_P^{(t)}(X;X') \triangleq D_P^{(t)}(P_{X,X'}||P_X \otimes P_{X'})$$
- The Chi-square information given by:
$$I_{\chi^2}(X;X') \triangleq \chi^2(P_{X,X'}||P_X \otimes P_{X'})$$
where $I_P^{(2)}(X;X') = I_{\chi^2}(X;X')$

III. Bounding Moments of Generalization Error

We begin by offering a general result inspired from [19] bounding the (absolute) expected value of an arbitrary function of the population and empirical risks under a joint measure in terms of the (absolute) expected value of the function of the population and empirical risks under the product measure.

Theorem 1. *Consider a measurable function $F(x,y) : \mathbb{R}^2 \to \mathbb{R}$. It follows that*
$$\begin{aligned}&\left|\mathbb{E}_{P_{W,S}}[F(L_P(W,\mu), L_E(W,S))]\right| \le \\ &\quad \mathbb{E}_{P_W \otimes P_S}[|F(L_P(W,\mu), L_E(W,S))|^q]^{\frac{1}{q}} (I_P^{(t)}(W;S)+1)^{\frac{1}{t}}\end{aligned} \tag{5}$$
where $t,q > 1$ such that $\frac{1}{t} + \frac{1}{q} = 1$, $P_S = \mu^{\otimes n}$ is the distribution of training set, and P_W and $P_{W,S}$ are the distribution of hypothesis and joint distribution of hypothesis and training set induced by learning algorithm $P_{W|S}$.

Proof. See Appendix A. □

Theorem 1 can now be immediately used to bound the moments of the generalization error of a randomized learning algorithm in terms of a power divergence, under the common assumption that the loss function is σ-subgaussian. [1] In the rest of paper, we assume that the loss function $\ell(w,z)$ is σ-subgaussian under distribution μ for all $w \in \mathcal{W}$.

Theorem 2. *The m-th moment of the generalization error of a randomized learning algorithm obeys the bound given by:*
$$\left|\overline{gen^m}(P_{W|S},\mu)\right| \le \sigma^m (\frac{mq}{n})^{\frac{m}{2}} e^{m/e} (I_P^{(t)}(W;S)+1)^{\frac{1}{t}} \tag{6}$$
provided that $t,q > 1$, $\frac{1}{t}+\frac{1}{q} = 1$, $mq > 2$ and $mq \in \mathbb{Z}^+$.

Proof. See Appendix B. □

Theorem 2 can also be immediately specialized to bound the moments of the generalization error of a randomized learning algorithm in terms of a chi-square divergence.

Corollary 1. *The m-th moment of the generalization error of a randomized learning algorithm obeys the bound given by:*
$$\left|\overline{gen^m}(P_{W|S},\mu)\right| \le \sigma^m (\frac{2m}{n})^{\frac{m}{2}} e^{m/e} \sqrt{I_{\chi^2}(W;S)+1} \tag{7}$$

Proof. This corollary follows immediately by setting $t = q = 2$ in Theorem 2. □

Interestingly, these moment bounds also appear to lead to a new average generalization error bound complementing existing ones in the literature.

Corollary 2. *The average generalization error can be bounded as follows:*
$$|\overline{gen}(P_{W|S},\mu)| \le \sigma\sqrt{\frac{q}{n}} e^{1/e} (I_P^{(t)}(W;S)+1)^{\frac{1}{t}} \tag{8}$$
provided that $q \ge 2$ for $q \in \mathbb{Z}^+$.

Proof. This corollary follows immediately by setting $m = 1$ in Theorem 2. □

Note that the chi-square information based expected generalization error upper bound is looser than the mutual information based counterpart in [3].

It is also interesting to reflect about how the generalization error moment bounds decay as a function of the training set size ingested by the learning algorithm. In general, information measures such as power information and chi-square information do not have to be finite, but these information measures can be shown to obey $0 \le I_P^{(t)}(W;S) \le R^t - 1$ and $0 \le I_{\chi^2}(W;S) \le R^2 - 1$, respectively, provided that [2]
$$\max_{(w,s)\in\mathcal{W}\times\mathcal{S}} \frac{P_{W|S}(w,s)}{P_W(w)} = R < \infty.$$
It follows immediately that the moments of the generalization error are governed by the upper bound given by:
$$\left|\overline{\text{gen}^m}(P_{W|S},\mu)\right| \le \sigma^m (\frac{2m}{n})^{\frac{m}{2}} e^{m/e} R \tag{9}$$
exhibiting a decay rate of the order $\mathcal{O}(\sqrt{\frac{1}{n^m}})$. Naturally, with the increase in the training set size, one would expect the empirical risk to concentrate around the population risk, and our bounds hint at the speed of such convergence.

It is also interesting to reflect about the tightness of the various generalization error moment bounds. In particular, in view of the fact that it may not be possible to compare directly information measures such as power information and chi-square information, the following Proposition puts forth conditions allowing one to compare the tightness of the bounds portrayed in Theorem 2 and Corollary 1 under the condition that the randomized learning algorithm ingests n i.i.d. input-output data examples.

Proposition 1. *The power information of order t generalization error m-th moment upper bound*
$$\left|\overline{gen^m}(P_{W|S},\mu)\right| \le \sigma^m (\frac{mq}{n})^{\frac{m}{2}} e^{m/e} (I_P^{(t)}(W;S)+1)^{\frac{1}{t}} \tag{10}$$
is looser than the chi-sqare information based bound
$$\left|\overline{gen^m}(P_{W|S},\mu)\right| \le \sigma^m (\frac{2m}{n})^{\frac{m}{2}} e^{m/e} \sqrt{I_{\chi^2}(W;S)+1} \tag{11}$$
provided that $(\frac{2(t-1)}{t})^{\frac{mt(t-1)}{(t-2)}} - 1 \le I_P^{(t)}(W;S)$ for $t > 2$ with $\frac{mt}{(t-1)} \in \mathbb{Z}^+$.

Proof. See Appendix C. □

[1] A random variable X is σ-subgaussian if $E[e^{\lambda(X-E[X])}] \le e^{\frac{\lambda^2\sigma^2}{2}}$ for all $\lambda \in \mathbb{R}$.

[2] This condition holds provided that $\mathcal{W} \times \mathcal{S}$ is countable.

For example, it turns out $1.34^{12} - 1 \leq I_P^{(3)}(W;S)$ guarantees a chi-square information based generalization error second moment bound to be tighter than the power information of order 3 based bound.

Finally, we offer an additional bound – applicable only to the second moment of the generalization error – leveraging an alternative proof route inspired by tools put forth in [5, Proposition 2]. It does not appear that [5, Proposition 2] can be used to generate generalization error higher-order moment bounds in closed form.

Theorem 3. *The second moment of the generalization error of a randomized learning algorithm can be bounded as follows:*

$$\overline{\mathit{gen}^2}(P_{W|S},\mu) \leq \frac{\sigma^2}{n}\left(16 I(W;S) + 9\right) \tag{12}$$

Proof. See Appendix D. □

The next proposition showcases that under certain conditions the mutual information based second moment bound can be tighter than the chi-square information bound.

Proposition 2. *The second moment of generalization error upper based on Chi-square information*

$$\overline{\mathit{gen}^2}(P_{W|S},\mu) \leq \frac{\sigma^2}{n} 4 e^{2/e}\sqrt{I_{\chi^2}(W;S)+1} \tag{13}$$

is looser than the upper bound based on mutual information in Theorem 3,

$$\overline{\mathit{gen}^2}(P_{W|S},\mu) \leq \frac{\sigma^2}{n}\left(16 I(W;S) + 9\right) \tag{14}$$

provided that $94 \leq I_{\chi^2}(W;S)$.

Proof. See Appendix E. □

IV. From Moments to High Probability Bounds

We now showcase how to use the moment upper bounds to bound the probability that the empirical risk deviates from the population risk by a certain amount, under a *single-draw* scenario where one draws a single hypothesis based on the training data [21].

Concretely, our following results leverage generalization error moment bounds to construct a generalization error high-probability bound. In particular, we offer a single draw upper bound on the generalization error by leveraging Theorem 2 in conjunction with Markov's inequality, that can be further optimized with respect to the moment order.

Theorem 4. *It follows that with probability at least* $1-\delta$ *for some* $\delta \in (0,1)$, $t > 1$, *by considering* $\beta = \frac{1}{t-1}\log\left(\frac{I_P^{(t)}(W;S)+1}{\delta^t}\right)$ *and under distribution* $P_{W,S}$ *the generalization error obeys:*

$$|\mathit{gen}(P_{W|S},\mu)| \leq \tag{15}$$
$$e^{1/e+1/2}\sqrt{\frac{2t\sigma^2}{n(t-1)}}\sqrt{\log\left(\sqrt[t]{I_P^{(t)}(W;S)+1}\right)+\log(\frac{1}{\delta})}$$

provided that $2 < \beta$ *and* $\beta \in \mathbb{Z}^+$.

Proof. See Appendix F. □

Remark 1. *The characterization in Theorem 4 can also be expressed in terms of* α*-Réyni divergence, by considering* $\beta = \frac{\alpha}{\alpha-1}\log(\frac{1}{\delta}) + D_\alpha(P_{W,S}||P_W \otimes P_S)$, *as follows:*

$$|\mathit{gen}(P_{W|S},\mu)| \leq \tag{16}$$
$$e^{1/e+1/2}\sqrt{\frac{2\sigma^2\left(D_\alpha(P_{W,S}||P_W\otimes P_S)+\log(\frac{1}{\delta})\right)}{n}}$$

provided that $2 < \beta$ *and* $\beta \in \mathbb{Z}^+$.

In [21, Corollary 4], a single draw upper bound is proposed which depends on two terms of α-Réyni divergences and the term $\frac{4\sigma^2}{n}\log(\frac{2}{\delta})$. Our upper bound, (16), depends on the α-Réyni divergence and also a smaller term $\frac{2\sigma^2}{n}\log(\frac{1}{\delta})$.

Corollary 3. *It follows that with probability at least* $1-\delta$ *for some* $\delta \in (0,1)$, *by considering* $\beta = \log\left(\frac{I_{\chi^2}(W;S)+1}{\delta^2}\right)$ *and under distribution* $P_{W,S}$ *the generalization error obeys:*

$$|\mathit{gen}(P_{W|S},\mu)| \leq \tag{17}$$
$$e^{1/e+1/2}2\sigma\sqrt{\frac{\log\left(\sqrt{I_{\chi^2}(W;S)+1}\right)+\log(\frac{1}{\delta})}{n}}$$

provided that $2 < \beta$ *and* $\beta \in \mathbb{Z}^+$.

Proof. This corollary follows immediately by setting $t = 2$ in Theorem 4. □

It is instructive to comment on how this information-theoretic based high-probability generalization error bound compares to other similar information-theoretic bounds such as in [3], [10] and [21]. Our single-draw bound dependence on δ (i.e. $\log(\frac{1}{\delta})$) is more beneficial than Xu *et al.*'s bound [3, Theorem 3] dependent on i.e. $(\frac{1}{\delta})$. Our single-draw bound based on chi-square information (along with bounds based on mutual information) is also typically tighter than maximal leakage based single draw bounds [10], [21].

A similar single-draw high probability upper bound based on chi-square information has also been provided in [10]. The approach pursued to lead to such bound in [10] is based on α-Réyni divergence and α-mutual information, whereas our approach leading to Corollary 3 is based on optimization of bounds to the moments of the generalization error with respect to order of the moments.

V. Numerical Example

We now illustrate our generalization error bounds within a very simple setting involving the estimation of the mean of a Gaussian random variable $Z \sim \mathcal{N}(\alpha,\beta^2)$ – where α corresponds to the (unknown) mean and β^2 corresponds to the (known) variance – based on n i.i.d. samples Z_i for $i = 1,\cdots,n$.

We consider the hypothesis corresponding to the empirical risk minimizer given by $W = \frac{Z_1+\cdots+Z_n}{n}$. We also consider the loss function given by

$$\ell(w,z) = \min((w-z)^2, c^2).$$

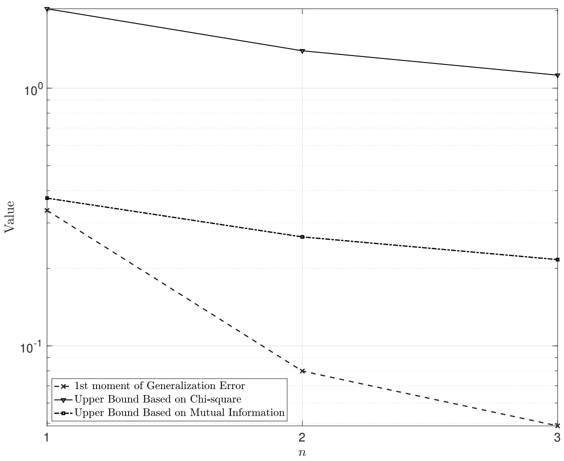

Fig. 1: First moment of the generalization error. The figure depicts the true values along with bounds based on mutual information and chi-square information.

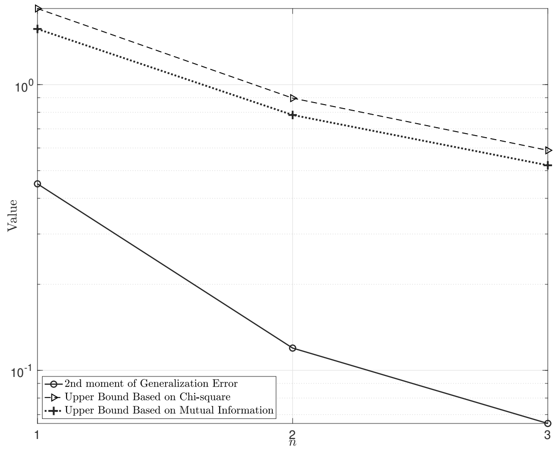

Fig. 2: Second moment of the generalization error. The figure depicts the true values along with bounds based on mutual information and chi-square information.

In view of the fact that the loss function is bounded within the interval $[0, c^2]$, it is also $\frac{c^2}{2}$-subgaussian so that we can apply the generalization error moments upper bounds offered earlier.

In our simulations, we consider $\alpha = 0$, $\beta^2 = 1$ and $c = \frac{2}{3}$. We compute the true generalization error numerically. We also compute chi-square and mutual information bounds to the moments of the generalization error appearing in Corollary 1 and Theorem 3. We focus exclusively on chi-square information – corresponding to power information of order 2 – because it has been established in Proposition 1 that the chi-square information bound can be tighter than the power information one under certain conditions. Both the chi-square information and the mutual information are evaluated numerically. Due to complexity in estimation of chi-square information and mutual information, we consider a relatively small number of training samples.

Fig.1 and Fig.2 demonstrate that the chi-square based bounds to the first and second moment of the generalization error is looser than the mutual information based bounds, as suggested earlier. Fig.3 also suggests that higher-order moments (and bounds) to the generalization error decay faster than lower-order ones, as highlighted earlier.

VI. Conclusion

We have introduced a new approach to obtain information-theoretic oriented bounds to the moments of generalization error associated with randomized supervised learning problems. We have discussed how these bounds relate to existing ones within the literature. Finally, we have also discussed how to leverage the generalization error moment bounds to derive a high probability bounds to the generalization error.

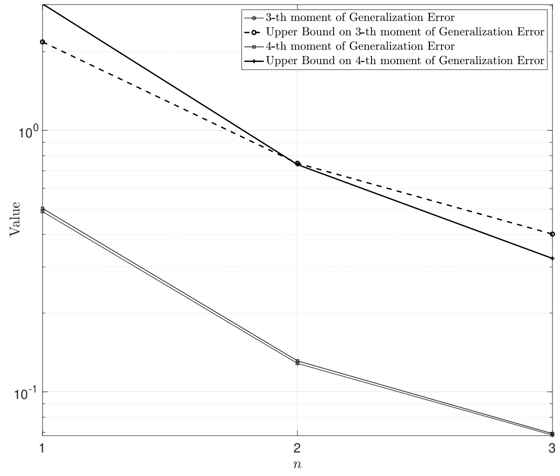

Fig. 3: Third and fourth moments of the generalization error. The figure depicts the true values along with bounds based on chi-square information.

References

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

[2] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*, vol. 1. MIT press Massachusetts, USA:, 2017.

[3] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, pp. 2524–2533, 2017.

[4] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *2016 IEEE Information Theory Workshop (ITW)*, pp. 26–30, IEEE, 2016.

[5] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, 2019.

[6] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020.

[7] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," in *Advances in Neural Information Processing Systems*, pp. 7234–7243, 2018.

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

[9] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for sgld via data-dependent estimates.," in *NeurIPS*, 2019.

[10] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via r\'enyi-, f-divergences and maximal leakage," *arXiv preprint arXiv:1912.01439*, 2019.

[11] G. Aminian, L. Toni, and M. R. Rodrigues, "Jensen-shannon information based characterization of the generalization error of learning algorithms," *2020 IEEE Information Theory Workshop (ITW)*, 2020.

[12] A. T. Lopez and V. Jog, "Generalization error bounds using wasserstein distances," in *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, IEEE, 2018.

[13] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via wasserstein distance," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 577–581, IEEE, 2019.

[14] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via wasserstein distance," *arXiv preprint arXiv:2101.09315*, 2021.

[15] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1475–1479, IEEE, 2017.

[16] D. A. McAllester, "Pac-bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[17] N. Thiemann, C. Igel, O. Wintenberger, and Y. Seldin, "A strongly quasi-convex pac-bayesian bound," in *International Conference on Algorithmic Learning Theory*, pp. 466–492, PMLR, 2017.

[18] O. Catoni, "A pac-bayesian approach to adaptive classification," *preprint*, vol. 840, 2003.

[19] P. Alquier and B. Guedj, "Simpler pac-bayesian bounds for hostile data," *Machine Learning*, vol. 107, no. 5, pp. 887–902, 2018.

[20] Y. Ohnishi and J. Honorio, "Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation," in *International Conference on Artificial Intelligence and Statistics*, pp. 1711–1719, PMLR, 2021.

[21] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *arXiv preprint arXiv:2005.08044*, 2020.

[22] P. Germain, A. Lacasse, F. Laviolette, M. March, and J.-F. Roy, "Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm," *Journal of Machine Learning Research*, vol. 16, no. 26, pp. 787–860, 2015.

[23] O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor, "Pac-bayes analysis beyond the usual bounds," *Advances in Neural Information Processing System*, 2020.

[24] A. Dhurandhar and A. Dobra, "Semi-analytical method for analyzing models and model selection measures based on moment analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, pp. 1–51, 2009.

[25] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy, "Sharper bounds for uniformly stable algorithms," in *Conference on Learning Theory*, pp. 610–626, PMLR, 2020.

[26] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[27] A. Guntuboyina, S. Saha, and G. Schiebinger, "Sharp inequalities for f-divergences," *IEEE transactions on information theory*, vol. 60, no. 1, pp. 104–121, 2013.

[28] P. P. Rigollet, "Lecture 2: Sub-gaussian random variables," in *high dimensional statistics—MIT Course No. 2.080J*, Cambridge MA: Massachusetts Institute of Technology, 2015. MIT OpenCourseWare.

[29] S. S. Dragomir and V. Gluscevic, "Some inequalities for the kullback-leibler and χ^2- distances in information theory and applications," *RGMIA research report collection*, vol. 3, no. 2, pp. 199–210, 2000.

Appendix A
Proof of Theorem 1

The result follows immediately by noting that:

$$|\mathbb{E}_{P_{W,S}}[F(L_P(w,\mu),L_E(w,S))]| \tag{18}$$

$$\leq \mathbb{E}_{P_{W,S}}[|F(L_P(w,\mu),L_E(w,S))|] \tag{19}$$

$$= \int |F(L_P(w,\mu),L_E(w,S))| \frac{dP_{W,S}}{d(P_W \otimes P_S)} d(P_W \otimes P_S) \tag{20}$$

$$\leq \mathbb{E}_{P_W \otimes P_S}[|F(L_P(w,\mu),L_E(w,S))|^q]^{\frac{1}{q}} \times \tag{21}$$

$$\left(\int (\frac{dP_{W,S}}{d(P_W \otimes P_S)})^t d(P_W \otimes P_S)\right)^{\frac{1}{t}}$$

$$= \mathbb{E}_{P_W \otimes P_S}[|F(L_P(w,\mu),L_E(w,S))|^q]^{\frac{1}{q}} (I_P^{(t)}(W;S)+1)^{\frac{1}{t}} \tag{22}$$

where (21) is due to Hölder inequality.

Appendix B
Proof of Theorem 2

This result follows from Theorem 1 by considering:

$$F(L_P(w,\mu),L_E(w,s)) = (L_P(w,\mu) - L_E(w,s))^m \tag{23}$$

We now have that:

$$\left|\overline{\text{gen}^m}(P_{W|S},\mu)\right| \leq \tag{24}$$

$$\mathbb{E}_{P_W}[\mathbb{E}_{P_S}[|(L_P(w,\mu) - L_E(w,s))|^{mq}]]^{\frac{1}{q}} (I_P^{(t)}(W;S)+1)^{\frac{1}{t}}$$

We also have that:

$$\mathbb{E}_{P_S}[|(L_P(w,\mu) - L_E(w,s))|^{mq}]^{\frac{1}{q}} \leq \sigma^m e^{m/e} (\frac{mq}{n})^{\frac{m}{2}} \tag{25}$$

in view of the fact that (a) the loss function is σ-subgaussian hence (b) $\text{gen}(P_{W|S},\mu)$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian and (c) [28, Lemma 1.4]. In [28, Lemma 1.4], it is assumed that (c) is valid for $mq > 2$ and $mq \in \mathbb{Z}^+$.This completes the proof.

Appendix C
Proof of Proposition1

This result follows from the inequality given by [27, Corollary 5.6]:

$$\sqrt{I_{\chi^2}(W;S)+1} \leq (I_P^{(t)}(W;S)+1)^{\frac{1}{2(t-1)}} \tag{26}$$

holding for $t > 2$. We then have that:

$$\sigma^m (\frac{2m}{n})^{\frac{m}{2}} e^{m/e} \sqrt{I_{\chi^2}(W;S)+1} \leq \tag{27}$$

$$\sigma^m (\frac{2m}{n})^{\frac{m}{2}} e^{m/e} (I_P^{(t)}(W;S)+1)^{\frac{1}{2(t-1)}} \leq$$

$$\sigma^m (\frac{mt}{n(t-1)})^{\frac{m}{2}} e^{m/e} (I_P^{(t)}(W;S)+1)^{\frac{1}{t}} \tag{28}$$

where the last inequality is valid if $(\frac{2(t-1)}{t})^{\frac{mt(t-1)}{(t-2)}} - 1 \leq I_P^{(t)}(W;S)$ for $t > 2$ and considering $\frac{mt}{(t-1)} \in \mathbb{Z}^+$.

APPENDIX D
PROOF OF THEOREM 3

The loss function is assumed to be σ-subgaussian under distribution μ for all $w \in \mathcal{W}$ hence – in view of the fact that the data samples are i.i.d. – $L_E(W,S)$ is $\frac{\sigma^2}{n}$-subgaussian and also gen$(P_{W|S},\mu)$ is $\frac{\sigma^2}{n}$-subgaussian under distribution P_S for all $w \in \mathcal{W}$.

It is possible to establish that the random variable $\text{gen}^2(P_{W|S},\mu) - \mathbb{E}_{P_S}[\text{gen}^2(P_{W|S},\mu)]$ is $(\frac{256\sigma^4}{n^2}, \frac{16\sigma^2}{n})$-subexponential [28, Lemma 1.12]. [3] It is worthwhile to mention that the random variable $\text{gen}^2(P_{W|S},\mu)$ is subexponential under distribution P_S for all $w \in \mathcal{W}$. We want to provide the upper bound on the expected value of $\text{gen}^2(P_{W|S},\mu)$ under the joint distribution $P_{W,S}$. Now, we have from the variational representation of the Kullback-Leibler distance that:

$$\lambda\mathbb{E}_{P_{W,S}}[\text{gen}^2(P_{W|S},\mu)] - \log(\mathbb{E}_{P_W\otimes P_S}[e^{\lambda\text{gen}^2(P_{W|S},\mu))}]) \tag{29}$$
$$\leq I(W;S)$$

As the $\text{gen}^2(P_{W|S},\mu)$ is $\frac{16\sigma^2}{n}$-subexponential under distribution μ for all $w \in \mathcal{W}$, we have:

$$\log(\mathbb{E}_{P_W\otimes P_S}[e^{\lambda(\text{gen}^2(P_{W|S},\mu)-\mathbb{E}_{P_S}[\text{gen}^2(P_{W|S},\mu)])}]) \leq \frac{128\lambda^2\sigma^4}{n^2} \quad \text{for} \quad |\lambda| \leq \frac{n}{16\sigma^2} \tag{30}$$

As $\text{gen}(P_{W|S},\mu)$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian, we also have $\mathbb{E}_{P_S}[\text{gen}^2(P_{W|S},\mu)] \leq \frac{\sigma^2}{n}$ for all $w \in \mathcal{W}$. Therefore the following inequality holds:

$$\log(\mathbb{E}_{P_W\otimes P_S}[e^{\lambda(\text{gen}^2(P_{W|S},\mu))}]) \leq \frac{128\lambda^2\sigma^4}{n^2} + \frac{\lambda\sigma^2}{n} \quad \text{for} \quad |\lambda| \leq \frac{n}{16\sigma^2} \tag{31}$$

This leads to the inequality:

$$\mathbb{E}_{P_{W,S}}[\text{gen}^2(P_{W|S},\mu)] \leq \frac{128\lambda\sigma^4}{n^2} + \frac{\sigma^2}{n} + \frac{I(W;S)}{\lambda} \tag{32}$$

holding for $|\lambda| \leq \frac{n}{16\sigma^2}$.

The final result follows by choosing $\lambda = \frac{n}{16\sigma^2}$.

APPENDIX E
PROOF OF PROPOSITION 2

The result follow from the inequality given by [29]:

$$I(W;S) \leq \log(I_{\chi^2}(W;S)+1) \tag{33}$$

We then have that:

$$16I(W;S) + 9 \leq 16\log(I_{\chi^2}(W;S)+1) + 9 \tag{34}$$

and, for $94 \leq I_{\chi^2}(W;S)$, we also have that:

$$16\log(I_{\chi^2}(W;S)+1) + 9 \leq 4e^{2/e}\sqrt{I_{\chi^2}(W;S)+1} \tag{35}$$

APPENDIX F
PROOF OF THEOREM 4

Consider that:

$$P_{W,S}(|\text{gen}(P_{W|S},\mu)| \geq r) = \tag{36}$$
$$P_{W,S}(|\text{gen}(P_{W|S},\mu)|^{2m} \geq r^{2m}) =$$
$$P_{W,S}(\text{gen}(P_{W|S},\mu)^{2m} \geq r^{2m}) \leq \tag{37}$$
$$\frac{\mathbb{E}_{P_{W,S}}[\text{gen}(P_{W|S},\mu)^{2m}]}{r^{2m}} \leq \tag{38}$$
$$\sigma^{2m}(\frac{2qm}{n})^m e^{2m/e}\frac{\sqrt[t]{(I_P^{(t)}(W;S)+1)}}{r^{2m}} \tag{39}$$

where the first inequality is due to Markov' inequality and the second inequality is due to Corollary 1. Consider also that

$$\delta = \sigma^{2m}\left(\frac{2qm}{n}\right)^m e^{2m/e}\frac{\sqrt[t]{(I_P^{(t)}(W;S)+1)}}{r^{2m}}$$

We then have immediately that with probability at least $1-\delta$ under the distribution $P_{W,S}$ it holds that:

$$|\text{gen}(P_{W|S},\mu)| \leq \tag{40}$$
$$\min_{m>0}\sigma\sqrt{\frac{2mq}{n}}e^{1/e}\frac{\sqrt[2mt]{I_P^{(t)}(W;S)+1}}{\sqrt[2m]{\delta}}$$

The value of m that optimizes the right hand side in the bound above is given by:

$$m^\star = \log\left(\frac{\sqrt[t]{I_P^{(t)}(W;S)+1}}{\delta}\right).$$

Based on (25), the $m^\star$ should also satisfy the conditions, $2 < \frac{m^\star t}{t-1}$ and $\frac{m^\star t}{t-1} \in \mathbb{Z}^+$. Therefore, we have $2 < \frac{1}{t-1}\log\left(\frac{I_P^{(t)}(W;S)+1}{\delta^t}\right)$and $\frac{1}{t-1}\log\left(\frac{I_P^{(t)}(W;S)+1}{\delta^t}\right) \in \mathbb{Z}^+$. The result then follows immediately by substituting $m^\star$ in (40).

[3] A random variable X is (σ^2,b)-subexponential if $\mathbb{E}_{P_X}[e^{\lambda(X-E[X])}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for all $|\lambda| \leq \frac{1}{b}$.