Report of Generalization Bounds via Convex Analysis

Yang Yiliu
1155157082

Abstract

The generalization error bound is always a hot topic in supervised machine learning. Several prior works have shown that this error can be bounded in terms of the mutual information between the algorithm input and output. In the paper titled "Generalization Bounds via Convex Analysis"[1], the authors propose a new bound in terms of general dependence measure. This report focuses on this paper and presents its main result and brief proof ideas. This report also discusses the tightness of key inequalities and the findings from this paper.

Note: Since this paper uses many tools from convex analysis, which is not covered in CSCI4230, readers are expected to read the textbook [2] in advance. In this report, readers are assumed to have basic knowledge of convex analysis.

1 Introduction

In the supervised machine learning field, generalization error bound is one of the key problems in validating the effectiveness of machine learning models. Prior works of [3] and [4] have shown that the expected generalization error of any algorithm $\mathcal{A}$ can be bounded in terms of Shannon's mutual information between the input dataset S with n data points and the output $W = \mathcal{A}(S)$. Suppose that the loss function $\ell(w, Z)$ of any fixed hypothesis w is σ-subgaussian, where Z is the random data point, then the generalization error is bounded in the following form:

$$|\mathbb{E}[\text{gen}(W, S)]| \leq \sqrt{\frac{\sigma^2 I(W; S)}{n}}, \tag{1}$$

where $I(\cdot; \cdot)$ is the mutual information function. This bound can be intuitively explained that the algorithm leak little information about the training data will have smaller generalization error.

However, the above bound has several weaknesses. For example, the mutual information $I(W; S)$ may be extremely large or even infinite when the algorithm significantly relies on the training data and leaks too much of them. One solution is to replace the input dataset S with a single data point Z_i [5], i.e., replace $I(W; S)$ with $I(W; Z_i)$. Recent works have also shown that other dependence measures such as Rényi's α- divergences and Csiszár's f-divergences can also be used to replace the mutual information to bound the generalization error [6]. Observing that there are many options to replace the mutual information, there is a need for a bounding form in terms of general dependence measures instead of only focusing on the mutual information.

Ideally, we may want to find a bound in the form of

$$|\mathbb{E}[\text{gen}(W, S)]| \leq \sqrt{\frac{CF(W; S)}{n}},$$

where $F(\cdot; \cdot)$ is one dependence measure. However, this inequality only holds when several assumptions are satisfied. In this paper, the authors model dependence measures as convex functions of the joint distribution of W_n and S_n with strongly convexity properties. Any such function H will satisfy a generalization bound of the form

$$|\mathbb{E}[\text{gen}(W_n, S_n)]| \leq \sqrt{\frac{C_{\ell,\mu,H} H(P_{W_n,S_n})}{n}}, \tag{2}$$

where P_{W_n,S_n} is the joint distribution of W_n and S_n, and $C_{\ell,\mu,H}$ is a constant depending on the loss function ℓ, the data distribution μ and the strong-convexity properties of H. Formal definitions of these terms will be introduced later.

The key contribution of this work is to propose a generalization error bound in terms of general dependence measures. Besides, the key idea to prove this bound is to use Fenchel-Young inequality with the Fenchel conjugate on the dependence measure H. Although this new bound does not indicate the generalization error decays faster than $n^{-1/2}$, it provides a general form of the constant factor.

2 Main result and proof ideas

2.1 Preliminaries

In the supervised learning scenario, we are given a dataset $S_n = \{Z_1, ..., Z_n\}$ from a distribution μ.[1] We are expected to propose a learning algorithm $\mathcal{A}$ that maps this dataset into a hypothesis $W_n = \mathcal{A}(S_n)$. We define the instance space $\mathcal{Z}$ and the hypothesis class $\mathcal{W}$. Thus, $Z_i \in \mathcal{Z}$, $S_n \in \mathcal{Z}^n$, and $W_n \in \mathcal{W}$.

We use loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$ to evaluate the performance of a hypothesis. Therefore, the training error is $L(W_n, S_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$ and the test error $\mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$. Overall, the generalization error is defined as

$$\text{gen}(W_n, S_n) = L(W_n, S_n) - \mathbb{E}_{Z\sim\mu}[\ell(W_n, Z)|W_n].$$

To illustrate the proof idea, we further define the following notations. We denote the set of all probability distributions over a given set $\mathcal{H}$ as $\mathcal{P}(\mathcal{H})$, and the set of bounded functions $\mathcal{H} \to \mathbb{R}$ as $\mathcal{F}(\mathcal{H})$. To simplify the writing, we use shorthand notation $\Delta = \mathcal{P}(\mathcal{W} \times \mathcal{S})$ and $\Gamma = \mathcal{P}(\mathcal{W})$. We denote the joint distribution of (W_n, S_n) by $P_n = P_{W_n,S_n}$, the marginal distribution of W_n by P_{W_n}. For any $P \in \Delta$, we use the notation $P_{|s} \in \Gamma$ to denote the conditional distribution of W given $S_n = s$. For any $g \in \mathcal{F}(\mathcal{W})$ and distribution $Q \in \Gamma$, we denote

$$\langle Q, g\rangle = \mathbb{E}_{W\sim Q}[g(W)].$$

Similarly, for any $f \in \mathcal{F}(\mathcal{W} \times \mathcal{S})$ and distribution $P \in \Delta$, we denote

$$\langle P, f\rangle = \mathbb{E}_{(W,S)\sim P}[f(W, S)] = \mathbb{E}_S[\langle P_{|S}, f(\cdot, S)\rangle].$$

For loss function, we define the centered loss $\bar{\ell}(w, z) = \ell(w, z) - \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$, the i-th sample loss as the function $\bar{\ell}_i(s, w) = \bar{\ell}(z_i, w)$ and the i-th partial average loss as $\bar{L}_i = \frac{1}{n}\sum_{j=1}^{i} \bar{\ell}_j$. Then the expected generalization error can be written as

$$\mathbb{E}[\text{gen}(W_n, S_n)] = \mathbb{E}[\bar{L}_n(W_n, S_n)] = \mathbb{E}_{(W,S)\sim P_n}[\bar{L}_n(W, S)] = \langle P_n, \bar{L}_n\rangle.$$

We define the dependence measure $H : \Delta \to \mathbb{R}_+$. We further define the conditional dependence measure $h : \Gamma \to \mathbb{R}_+$, i.e., $H(P) = \mathbb{E}_S[h(P_{|S})]$. We assume h is convex and lower semicontinuous on Γ and satisfies $h(P_{W_n}) = 0$.

For any $Q, Q' \in \Gamma$, we define subdifferential of h at Q' denoted by $\partial h(Q')$ as the set of functions $g \in \mathcal{F}(\mathcal{W})$ such that

[1] This section mainly follows the original paper.

$$h(Q) \geq h(Q') + \langle Q - Q', g \rangle$$

holds. Furthermore, we say a conditional dependence measure h is α-strongly convex with respect to a norm $||\cdot||$ if for any $Q, Q' \in \Gamma$ and $g \in \partial h(Q')$,

$$h(Q) \geq h(Q') + \langle Q - Q', g \rangle + \frac{\alpha}{2}||Q - Q'||^2 \tag{3}$$

holds. For any norm $||\cdot||$, we define the associated dual norm as

$$||f||_* = \sup_{Q,Q' \in \Gamma : ||Q-Q'|| \leq 1} \langle Q - Q', f \rangle$$

for any bounded function $f \in \mathcal{F}(\mathcal{W})$. Apart from the defined training set $S_n = \{Z_i\}_{i=1}^n$, we define the independent "ghost dataset" $S'_n = \{Z'_i\}_{i=1}^n$ consisting of i.i.d. samples from the same distribution μ. For $i \in [n]$, we define the "mixed bag" dataset $S_n^{(i)} = \{Z_1, Z_2, ..., Z_i, Z'_{i+1}, ..., Z'_n\}$ and the corresponding output $W^{(i)} = \mathcal{A}(S_n^{(i)})$. We define P_i as the joint distribution of $(W^{(i)}, S_n)$. Finally, for any i, we define Δ_i as the convex hull of all distributions $\{P_k\}_{k=0}^i : \Delta_i = \{P \in \Delta : \sum_{k=0}^i \alpha_k P_k, \alpha_k \geq 0, \sum_{k=0}^i \alpha_k = 1\}$.

2.2 Main result

The main result follows the form of equation 2. The constant $C_{\ell,\mu,H}$ is replaced by $\frac{4\mathbb{E}[||\bar{\ell}(\cdot,Z)||_*^2]}{\alpha}$.

Theorem 1. *Given a dependence measure H, where h is α-strongly convex with respect to the norm $||\cdot||$. With the help of the dual norm $||\cdot||_*$, the generalization error of $\mathcal{A}$ is bounded as*

$$|\mathbb{E}[gen(W_n, S_n)]| \leq \sqrt{\frac{4H(P_n)\mathbb{E}[||\bar{\ell}(\cdot,Z)||_*^2]}{\alpha n}}.$$

2.3 Proof ideas

The proof consists of one equation (5), one theorem (2), and one lemma (3). Because the whole proof is too long and tedious, to make this report concise and easy to understand, I will only show the basic proof ideas of these three items.

Recall that the proof will mainly use Fenchel-Young inequality. We define the Fenchel conjugate of dependence measure H as the potential function Φ that maps functions $f : \mathcal{W} \times \mathcal{S} \to \mathbb{R}$ to reals. Φ is defined as

$$\Phi(f) = \sup_{P \in \Delta_n} \{\langle P, f \rangle - H(P)\}. \tag{4}$$

This potential function can also be interpreted as the "overfitting potential". Noting that this potential function Φ is exactly the Fenchel conjugate of H, then apply the Fenchel-Young inequality to the expected generalization error as follows:

$$\eta\mathbb{E}[\text{gen}(W_n, S_n)] = \eta\langle P_n, \bar{L}_n \rangle \leq H(P_n) + \Phi(\eta\bar{L}_n). \tag{5}$$

Here we already get a bound on the expected generalization error. We will further get a bound on the potential function $\Phi(\eta\bar{L}_n)$. Noting that $\mathbb{E}[\text{gen}(W_n, S_n)] \leq \frac{H(P_n)}{\eta} + \frac{\Phi(\eta\bar{L}_n)}{\eta} \leq \sqrt{\frac{4H(P_n)\Phi(\eta\bar{L}_n))}{\eta^2}}$, we will expect to get a bound on $\Phi(\eta\bar{L}_n)$ of the order $\frac{\eta^2}{n}$.

We further define the subdifferential of Φ as the convex hull of the maximizers of $\{\langle P, f \rangle - H(P)\}$:

$$\partial\Phi(f) = \text{conv}\left(\operatorname*{argmax}_{P \in \Delta_n}\{\langle P, f \rangle - H(P)\}\right).$$

Thus, we can define the corresponding generalized Bregman divergence as

$$\mathcal{B}_\Phi(g||f) = \Phi(g) - \Phi(f) + \sup_{P \in \partial\Phi(f)} \langle P, f - g \rangle.$$

Then we can have the following theorem to bound the potential function Φ and convert it into the Bregman divergence $\mathcal{B}_\Phi$.

Theorem 2. *For any $\eta \in \mathbb{R}$, the overfitting potential satisfies*

$$\Phi(\eta \bar{L}_n) \leq \sum_{i=1}^{n} \mathcal{B}_\Phi(\eta \bar{L}_i || \eta \bar{L}_{i-1}).$$

Note that $\Phi(0) = 0$ due to $H(P_0) = 0$. Starting from the left hand side,

$$\begin{aligned}
\Phi(\eta \bar{L}_n) &= \sum_{i=1}^{n} \left(\Phi(\eta \bar{L}_i) - \Phi(\eta \bar{L}_{i-1}) \right) + \Phi(0) \\
&= \sum_{i=1}^{n} \left(\mathcal{B}_\Phi(\eta \bar{L}_i || \eta \bar{L}_{i-1}) - \eta \sup_{P \in \partial\Phi(\eta \bar{L}_{i-1})} \langle P, \bar{L}_{i-1} - \bar{L}_i \rangle \right) \\
&= \sum_{i=1}^{n} \left(\mathcal{B}_\Phi(\eta \bar{L}_i || \eta \bar{L}_{i-1}) + \frac{\eta}{n} \inf_{P \in \partial\Phi(\eta \bar{L}_{i-1})} \langle P, \bar{\ell}_i \rangle \right)
\end{aligned}$$

Then we have to show $\inf_{P \in \partial\Phi(\eta \bar{L}_{i-1})} \langle P, \bar{\ell}_i \rangle \leq 0$. I will not cover the whole proof but give you a brief idea.

As $Z_1, ..., Z_{i-1}$ are independent of Z_i, seeing the former or not will affect the helpfulness of Z_i. Therefore, for any $P \in \text{argmax}_{P \in \Delta_n} \{ \langle P, \eta \bar{L}_{i-1} \rangle - H(P) \}$, which means P overfits the first $i-1$ data points, $\langle P, \bar{\ell}_i \rangle = 0$. As these P's are in $\partial\Phi(\eta \bar{L}_{i-1})$, $\inf_{P \in \partial\Phi(\eta \bar{L}_{i-1})} \langle P, \bar{\ell}_i \rangle \leq 0$ holds.

So far, we have gained a bound of the order n (contributed by the summation). We want to derive a new bound on this Bregman divergence of the order $\frac{\eta^2}{n^2}$. Then we have the following lemma.

Lemma 3. *Suppose that h is α-strongly convex with respect to the norm $|| \cdot ||_*$. Then,*

$$\mathcal{B}_\Phi \left(\eta \bar{L}_i || \eta \bar{L}_{i-1} \right) \leq \frac{\eta^2 \mathbb{E}_Z \left[|| \bar{\ell}(\cdot, Z) ||_*^2 \right]}{\alpha n^2}$$

I will also not give detailed proof. The brief idea is due to the duality of strongly convexity and smoothness, the α-strongly convexity of H implies $\frac{1}{\alpha}$-strongly convexity smoothness of its Fenchel conjugate Φ. Thus, by the definition of Bregman divergence, we can have

$$\mathcal{B}_\Phi \left(\eta \bar{L}_i || \eta \bar{L}_{i-1} \right) \leq \frac{1}{\alpha} || \eta \bar{L}_i - \eta \bar{L}_{i-1} ||_{\mu,*}^2 = \frac{\eta^2 \mathbb{E}_Z \left[|| \bar{\ell}(\cdot, Z) ||_*^2 \right]}{\alpha n^2}.$$

By combining equation 5, Theorem 2 and Lemma 3, we can have

$$\eta \langle P_n, \bar{L}_n \rangle \leq H(P_n) + \frac{\eta^2 \mathbb{E}_Z \left[|| \bar{\ell}(\cdot, Z) ||_*^2 \right]}{\alpha n}.$$

Considering both $\eta > 0$ and $\eta < 0$, we have

$$|\mathbb{E}[\text{gen}(W_n, S_n)]| = |\langle P_n, \bar{L}_n \rangle| \leq |\frac{H(P_n)}{\eta} + \frac{\eta \mathbb{E}_Z \left[|| \bar{\ell}(\cdot, Z) ||_*^2 \right]}{\alpha n}| \leq \sqrt{\frac{4H(P_n) \mathbb{E}[|| \bar{\ell}(\cdot, Z) ||_*^2]}{\alpha n}}. \quad (6)$$

3 Application in KL-divergence

We can choose KL-divergence as the dependence measure, i.e.,

$$h(Q) = \mathcal{D}_{\text{KL}}(Q||Q_0) = \int_{\mathcal{W}} \log \frac{dQ}{dQ_0} dQ_0,$$

where Q and Q_0 are the marginal hypothesis distribution of P and P_0. This KL-divergence is 1-strongly convex with respect to the total variation distance $||Q - Q'||_{TV} = \sup_{f:||f||_\infty \leq 1} \langle f, Q - Q' \rangle$ and its dual norm $||f||_\infty = \sup_{w \in \mathcal{W}} |f(w)|$. With this marginal h, the dependence measure $H(P) = \mathbb{E}[\mathcal{D}_{\text{KL}}(P_{|S}||Q_0)] = \mathcal{D}_{\text{KL}}(P||P_0)$. Then apply Theorem 1, we can get

Corollary 1. *The generalization error of any learning algorithm satisfies*

$$|\mathbb{E}[gen(W_n, S_n)]| \leq \sqrt{\frac{4\mathcal{D}_{KL}(P_n||P_0)\mathbb{E}_Z[||\bar{\ell}(\cdot, Z)||_\infty^2]}{n}}$$

This paper also mentions other applications in p-norm divergences and smoothed relative entropy, and shows that Theorem 1 yields meaningful results based on these dependence measures.

4 Discussions

4.1 Inequality tightness

I will discuss the tightness of key inequalities used in this proof in the following paragraphs.

Equation 5 The inequality is tight. As suggested by [7], for all $P_n \in \partial\Phi(\eta \bar{L}_n)$, $\eta\langle P_n, \bar{L}_n \rangle = H(P_n) + \Phi(\eta \bar{L}_n)$, where $\partial\Phi(\eta \bar{L}_n) \neq \phi$.

Theorem 2 The inequality is tight. The key factor of this inequality is $\inf_{P \in \partial(\eta \bar{L}_{i-1})} \langle P, \bar{\ell}_i \rangle$. This term has a possibility of being negative only when $\text{argmax}_{P \in \Delta_n} \{\langle P, \eta \bar{L}_{i-1} \rangle - H(P)\}$ is non-convex. Therefore, any loss function and dependence measure yielding convex solution space will have equality in this theorem. For example, when $\ell = 0$, the solution space will be convex because H is a convex function by definition.

Lemma 3 The inequality is tight. As [8] suggests, when the conditional dependence measure H is exactly α-strongly convex everywhere, i.e., equation 3 takes equality for all Q and Q', its Fenchel conjugate Φ will also be exactly $\frac{1}{\alpha}$-strongly smooth everywhere. Therefore, this inequality is tight.

Equation 6 The inequality is tight. When $\eta = \sqrt{\frac{\alpha n H(P_n)}{\mathbb{E}_Z[||\bar{\ell}(\cdot, Z)||_*^2]}}$, the equality holds.

Overall, these four key inequalities in this paper are tight. However, if more assumptions are made on dependence measures, the result may be tighter.

4.2 Insight into learning algorithms

Based on my current shallow understanding of machine learning, basic generalization error analysis focuses on the hypothesis class $\mathcal{W}$ and VC Dimension d_{VC} of $\mathcal{W}$. For example, for classification models with finite hypothesis class, with probability at least $1 - \delta$,

$$gen(\mathcal{W}) \leq \sqrt{\frac{1}{2n} \log \frac{2|\mathcal{W}|}{\delta}}.$$

For infinite hypothesis classes, we can use VC Dimension to analyze the generalization error. With probability at least $1 - \delta$,

$$gen(\mathcal{W}) \leq \sqrt{\frac{8}{n} \log \frac{4((2n)^{d_{\text{VC}}} + 1}{\delta}}.$$

However, these bounds only consider the hypothesis class but neither the data distribution nor the learning algorithm. This paper gives a perspective of viewing the generalization error in terms of the learning algorithm $\mathcal{A}$, which is reflected by the dependence degree of the training dataset and the algorithm output. This paper further generalizes the dependence measure from specific functions (e.g., Shannon's mutual information) to general strongly convex functions.

This result gives us an insight into the effect of learning algorithms in generalization error bounding. Nowadays, there are only several kinds of popular learning algorithms in modern machine learning models, such as gradient descent on neural networks. We can narrow the analysis target to specific learning algorithms to make appropriate strong assumptions and gain tighter bounds.

4.3 Miscellaneous

Order analysis As suggested by many papers, the decay rate of expected generalization error is expected to be $n^{-1/2}$. Therefore, the proposed bound should also have this order. When deriving the bound, we should always keep in mind what order we currently have and what order we expect to get to have the final order $n^{-1/2}$. For example, we expect to get a bound on $\Phi(\eta \bar{L}_n)$ of the order $\frac{\eta^2}{n}$ in equation 5, and a bound on $\mathcal{B}_\Phi(\eta \bar{L}_i || \eta \bar{L}_{i-1})$ of the order $\frac{\eta^2}{n^2}$ in Theorem 2. This technique is helpful when deriving an inequality where the final order is known or strongly believed. Besides, this method is similar to Dimensional analysis, a well-known method in physical analysis.

Probability bounds Referring to the discussion in 4.2, the generalization error bounds in terms of the hypothesis class $\mathcal{H}$ have possibility guarantees, i.e., the inequalities hold with high possibilities. However, the proposed bound only bounds the expected value. It is unknown if we can have a tighter bound on the exact error value with a high probability.

From specific to general This paper is inspired by all previous studies on the expected generalization error in terms of Shannon's mutual information, "single-letter" mutual information, Rényi's α-divergences and Csiszár's f-divergences and many other studies. And then the author can propose a bound in terms of general dependence measures. This gives us a point that we can initially focus on specific scenarios or strong assumptions. When satisfactory results are gained, we can then put them into general.

5 Conclusion

This report shows the main result of the original paper [1] with brief ideas of the proof. This report also discusses the tightness of key inequalities and some findings from the paper. Briefly, this paper proposes a bound on the expected generalization error in terms of general dependence measure, which allows researchers to gain the upper bound using any dependence measure that satisfies the requirements. In the future, researchers may find better dependence measures to capture the dependence between the input dataset and output hypothesis and get a tighter bound on expected generalization error accordingly compared with Shannon's mutual information. Researchers may also make stronger assumptions on the dependence measure to gain tighter bounds.

References

[1] Gábor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In *Conference on Learning Theory*, pages 3524–3546. PMLR, 2022.

[2] Constantin Zalinescu. *Convex analysis in general vector spaces*. World scientific, 2002.

[3] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

[4] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

[5] Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, 2020.

[6] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8): 4986–5004, 2021.

[7] f10w. How equality in fenchel-young inequality characterizes subdifferential?, May 2021. URL `https://math.stackexchange.com/questions/1427975/how-equality-in-fenchel-young-inequality-characterizes-subdifferential`.

[8] user1736. Strong convexity and strong smoothness duality, Jan 2015. URL `https://math.stackexchange.com/questions/1279653/strong-convexity-and-strong-smoothness-duality`.