A short note on the information bounds of generalization errors

Yen-Chi Chen
University of Washington
February 14, 2022

References:

- Xu, A., & Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. Advances in Neural Information Processing Systems, 30.
- Russo, D., & Zou, J. (2019). How much does your data exploration overfit? Controlling bias via information usage. IEEE Transactions on Information Theory, 66(1), 302-323.
- Neu, G., Lugosi, G. (2022). Generalization Bounds via Convex Analysis. arXiv:2202.04985

1 Problem setup

In this note, I will summarize a simple information bound on the generalization error. Consider a classical prediction problem where our training data is

$$(X_1,Y_1),\cdots,(X_n,Y_n)\in\mathcal{X}\times\mathcal{Y}$$

that are IID from some unknown distribution P_{XY}. For simplicity, we may denote $Z_i=(X_i,Y_i)$ so that the training data can be viewed as IID from P_Z. We denote $P_Z^{\otimes n}=P_Z\times P_Z\times\cdots P_Z$ as the joint PDF of $(Z_1,\cdots,Z_n)$.

In a typical supervised learning, we try to construct a predictor $c:\mathcal{X}\to\mathcal{Y}$. To simplify the problem, we assume that this predictor is indexed by a parameter θ, so we can write $c(x)=c_\theta(x)$.

Let $\ell_0:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ be the loss function. For a new observation $(X',Y')=Z'$ and a given predictor c_θ, the loss incurred is

$$\ell_0(c_\theta(X'),Y')=\ell(\theta,Z').$$

Namely, *we can rewrite the loss in terms of* θ *and* Z. This expression will be a key in our future analysis.

With the above notations, we define both the training and test error for a given classifier c_θ:

- Training error (empirical risk):

$$\widehat{R}_n(\theta)=\frac{1}{n}\sum_{i=1}^n\ell_0(c_\theta(X_i),Y_i)=\frac{1}{n}\sum_{i=1}^n\ell(\theta,Z_i).$$

  If the predictor is trained from the training data, we plug-in $\widehat{\theta}=\widehat{\theta}(Z_1,\cdots,Z_n)$ into the above expression and obtain

$$\widehat{R}_n(\widehat{\theta})=\frac{1}{n}\sum_{i=1}^n\ell(\widehat{\theta},Z_i).$$

- Test error (test risk):
$$R(\theta) = \mathbb{E}[\ell_0(c_\theta(X'), Y')] = \mathbb{E}[\ell(\theta, Z')].$$
When the predictor is $\widehat{\theta}$, its test risk is
$$R(\widehat{\theta}) = \mathbb{E}[\ell(\widehat{\theta}, Z')|\widehat{\theta}].$$
The expectation only applies to Z', not $\widehat{\theta}$.

- Generalization error (generalization risk): In this case, the generalization error is the expected difference between the training error and test error of the estimator $\widehat{\theta}$, which is
$$\mathsf{Gen} = \mathbb{E}[\widehat{R}_n(\widehat{\theta}) - R(\widehat{\theta})] = \mathbb{E}[\widehat{R}_n(\widehat{\theta})] - \mathbb{E}[R(\widehat{\theta})]. \tag{1}$$
The expectation is applied to the training data $Z_1, \cdots, Z_n$, which includes $\widehat{\theta}$.

  In the end, we will show that
$$\mathsf{Gen} = \mathbb{E}[\widehat{R}_n(\widehat{\theta}) - R(\widehat{\theta})] \le O\left(\sqrt{I(\widehat{\theta}, Z_1, \cdots, Z_n)}\right),$$
where $I(\widehat{\theta}, Z_1, \cdots, Z_n)$ is the mutual information between $\widehat{\theta}$ and the training data $(Z_1, \cdots, Z_n)$.

2 Generalization error and independence

A key insight is that the generalization error in equation (1) is related to the difference between independent and dependent distributions. Recall that $P_{Z,n}$ is the joint distribution of $Z_1, \cdots, Z_n$. We denote $P_{\theta, Z_1, \cdots, Z_n}$ to be the joint distribution of $\widehat{\theta}, Z_1, \cdots, Z_n$ and $P_{\theta|Z,n}$ to be the conditional distribution of $\widehat{\theta}|Z_1, \cdots, Z_n$. Note that here we assume that $\widehat{\theta}$ is a randomized estimator so that even if the data is held fixed, $\widehat{\theta}$ may still be random.

We can rewrite $\mathbb{E}[\widehat{R}_n(\widehat{\theta})]$ as
$$\begin{aligned}
\mathbb{E}[\widehat{R}_n(\widehat{\theta})] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell(\widehat{\theta}, Z_i)\right] \\
&= \int \frac{1}{n}\sum_{i=1}^n \ell(\theta, z_i) P_{\theta, Z_1, \cdots, Z_n}(d\theta, dz_1, \cdots, dz_n) \\
&= \int \frac{1}{n}\sum_{i=1}^n \ell(\theta, z_i) P_{\theta|Z,n}(d\theta) \prod_{j=1}^n P_Z(dz_j)
\end{aligned} \tag{2}$$

Thus, in the expectation of the empirical risk, we are integrating over the *joint distribution*
$$P_{\theta, Z_1, \cdots, Z_n} = P_{\theta|Z,n} \cdot P_Z^{\otimes n}, \tag{3}$$
where $P_Z^{\otimes n} = P_Z \times P_Z \times \cdots P_Z$ is the joint distribution of $Z_1, \cdots, Z_n$.

Now we turn to our analysis on the test risk. Let P_θ be the marginal distribution of $\widehat{\theta}$ from $P_{\theta,Z_1,\cdots,Z_n}$. We can rewrite the test risk as

$$
\begin{aligned}
\mathbb{E}[R(\widehat{\theta})] = \mathbb{E}\left[\ell(\widehat{\theta},Z')\right]& \\
&\int \ell(\theta,z')P_\theta(d\theta)\cdot P_Z(dz') \\
&\int \frac{1}{n}\sum_{i=1}^{n}\ell(\theta,z_i')P_\theta(d\theta)\cdot P_Z(dz_i') \\
&\int \frac{1}{n}\sum_{i=1}^{n}\ell(\theta,z_i')P_\theta(d\theta)\cdot \prod_{j=1}^{n}P_Z(dz_j').
\end{aligned}
\tag{4}
$$

So in the test risk, we are integrating over the joint distribution

$$
P_{\theta,Z_1',\cdots,Z_n'} = P_\theta \cdot P_Z^{\otimes n}, \tag{5}
$$

which is the case of *assuming* θ *and* $Z_1',\cdots,Z_n'$ *are independent!*

As a result, the generalization error is the difference between expectation of dependent $\theta, Z_1,\cdots,Z_n$ and the independent θ and $Z_1,\cdots,Z_n$. From this perspective, you can see why the information bounds on dependence will be useful in controlling the generalization errors.

3 A useful mutual information bound

From the above analysis, we have seen that we may bound the generalization errors using measures of dependency. Here is a simple bound based on mutual information.

Lemma 1 *Let (U,V) be two continuous random vectors that are dependent with each other. Let $(\bar{U},\bar{V})$ be random vectors such that $\bar{U} \stackrel{d}{=} U$ and $\bar{V} \stackrel{d}{=} V$ with $\bar{U} \perp \bar{V}$. Namely, $\bar{U}$ has the same distribution as U but it is independent of $\bar{V}$. Consider a function $f(u,v)$. If $f(\bar{U},\bar{V})$ is $\sigma-$sub-Gaussian, then*

$$
|\mathbb{E}[f(U,V)] - \mathbb{E}(f(\bar{U},\bar{V}))| \leq \sqrt{2\sigma^2 I(U,V)},
$$

where $I(U,V)$ is the mutual information between U and V.

Proof. Recall that the mutual information $I(U,V) = KL(p_{U,V}||p_U \cdot p_V)$, where KL is the Kullback-Leiber divergence and $p_{U,V}$ is the joint PDF of U,V.

A key of the proof is the following variational form of the KL-divergence. For any two PDF q,π,

$$
KL(q||\pi) = \sup_\eta \left\{\int \eta(x)dq(x) - \log\int e^{\eta(x)d\pi(x)}\right\};
$$

see, e.g. Corollary 4.15 of

S. Boucheron, G. Lugosi, and P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford Univ. Press, 2013.

We choose q to be the PDF of (U,V) and π to be the PDF of $(\bar{U},\bar{V})$ and $\eta = \lambda \cdot f$, where λ is a free parameter. Then the above variational form implies

$$\begin{aligned}
I(U,V) &= KL(p_{U,V}||p_U \cdot p_V) \\
&\geq \int \lambda f(u,v) dp_{U,V}(u,v) - \log \int e^{\lambda f(u,v)} dp_U(u) dp_V(v) \\
&= \mathbb{E}(\lambda f(U,V)) - \log \mathbb{E}\left(e^{\lambda f(\bar{U},\bar{V})}\right) \\
&= \mathbb{E}(\lambda f(U,V)) - \log \mathbb{E}\left(e^{\lambda[f(\bar{U},\bar{V}) - \mathbb{E}(f(\bar{U},\bar{V}))]}\right) - \mathbb{E}(\lambda f(\bar{U},\bar{V})).
\end{aligned}$$

By σ-sub-Gaussian property of $f(\bar{U},\bar{V})$, we have

$$\log \mathbb{E}\left(e^{\lambda[f(\bar{U},\bar{V}) - \mathbb{E}(f(\bar{U},\bar{V}))]}\right) \leq \frac{1}{2}\lambda^2\sigma^2.$$

So the above inequality becomes

$$\begin{aligned}
I(U,V) &\geq \mathbb{E}(\lambda f(U,V)) - \log \mathbb{E}\left(e^{\lambda[f(\bar{U},\bar{V}) - \mathbb{E}(f(\bar{U},\bar{V}))]}\right) - \mathbb{E}(\lambda f(\bar{U},\bar{V})) \\
&\geq \lambda \mathbb{E}(f(U,V)) - f(\bar{U},\bar{V})) - \frac{1}{2}\lambda^2\sigma^2 \\
&\geq \frac{\mathbb{E}^2(f(U,V)) - f(\bar{U},\bar{V}))}{2\sigma^2},
\end{aligned}$$

where the last inequality follows from optimizing λ, which occurs at $\lambda^* = \frac{\mathbb{E}(f(U,V)) - f(\bar{U},\bar{V}))}{\sigma^2}$.

Thus, this implies

$$|\mathbb{E}[f(U,V)] - \mathbb{E}(f(\bar{U},\bar{V}))| \leq \sqrt{2\sigma^2 I(U,V)},$$

which completes the proof.

□

4 Conclusion

By applying Lemma 1 with $\widehat{\theta} = U$ and $(Z_1, \cdots, Z_n) = V$, we conclude that

$$\mathsf{Gen} = \mathbb{E}[\widehat{R}_n(\widehat{\theta}) - R(\widehat{\theta})] \leq \sqrt{2\sigma^2 I(\widehat{\theta}, Z_1, \cdots, Z_n)}$$

assuming the σ-sub-Gaussianity.