An Information-Theoretic Route from Generalization in Expectation to Generalization in Probability (Supplementary File)

Ibrahim Alabdulmohsin
Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

1 Proof of Proposition 1

Let $\mathcal{Z} = [0,1]$ be an instance space with a continuous marginal density $p(z)$ (hence, has no atoms) and let $\mathcal{Y} = \{-1,+1\}$ be the target set. Let $h^\star : \mathcal{Z} \to \{-1,+1\}$ be some *fixed* predictor, such that $p\{h^\star(Z) = 1\} = \frac{1}{2}$, where the probability is evaluated over the random choice of $Z \in \mathcal{Z}$. In other words, the marginal distribution of the labels predicted by $h^\star(\cdot)$ is uniform[1].

Next, let the hypothesis space $\mathcal{H}$ be the set of predictors from $\mathcal{Z}$ to $\{-1,+1\}$ that output a label in $\{-1,+1\}$ uniformly at random everywhere in $\mathcal{Z}$ except at a finite number of points. Therefore, the hypothesis $H : \mathcal{Z} \to \{-1,+1\}$ selected by the learning algorithm is a predictor. Define the parametric loss by $L(Z;H) = \mathbb{I}\big\{H(Z) \neq h^\star(Z)\big\}$.

Next, we construct a learning algorithm $\mathcal{L}$ that generalizes perfectly in expectation but it does not generalize in probability. The learning algorithm $\mathcal{L}$ simply picks one of $H^0_{S_m}(\cdot)$ or $H^1_{S_m}(\cdot)$ with equal probability, where:

$$H^0_{S_m}(Z) = \begin{cases} -h^\star(Z) & \text{if } Z \in S_m \\ \text{Uniform}(-1,+1) & \text{if } Z \notin S_m \end{cases}$$

$$H^1_{S_m}(Z) = \begin{cases} h^\star(Z) & \text{if } Z \in S_m \\ \text{Uniform}(-1,+1) & \text{if } Z \notin S_m \end{cases}$$

Because $\mathcal{Z}$ is uncountable, where the probability of seeing the same observation Z twice is zero, $R_{true}(H) = \frac{1}{2}$ for this learning algorithm. Thus:

$$R_{gen}(\mathcal{L}) = \mathbb{E}_{S_m,H}\big[R_{emp}(H;S_m) - R_{true}(H)\big] = 0$$

However, the empirical risk for any S_m satisfies $R_{emp}(H;S_m) \in \{0,1\}$ while the true risk always satisfies $R_{true}(H) = \frac{1}{2}$, as mentioned earlier. Hence, the statement of the proposition follows.

[1]These assumptions are satisfied, for example, if $p(z)$ is uniform in $[0,1]$ and $h^\star(Z) = \mathbb{I}\{Z < \frac{1}{2}\}$.

Finally, we prove that the algorithm does not generalize uniformly in expectation. There are, at least, two ways of showing this. The first approach is to use the equivalence between uniform generalization and algorithmic stability as stated in Theorem 1. Given the hypothesis $H \in \{H^0_{S_m},\ H^1_{S_m}\}$ learned by the algorithm constructed here, the marginal distribution of an individual training example $p(Z_{trn}|H)$ is uniform over the sample S_m. This follows from the fact that the hypothesis H has to encode the entire sample S_m. However, the probability of seeing the same observation twice is zero (by construction). Hence, $||p(Z_{trn}),\ p(Z_{trn}|H)||_{\mathcal{T}} = 1$ for all H. This shows that $\mathbb{S}(\mathcal{L}) = 0$ for all $m \geq 1$, and the learning algorithm is not stable. Therefore, by Theorem 1, it does not generalize uniformly. Note that we used the information-theoretic interpretation of uniform generalization.

The second approach is to use the statistical interpretation of uniform generalization. Let $H \in \{H^0_{S_m},\ H^1_{S_m}\}$ be the hypothesis inferred by the learning algorithm above, and consider the following *different* parametric loss:

$$L(Z;H^k_{S_m}) = \mathbb{I}\big\{(-1)^{k+1}H^k_{S_m}(Z) \neq h^\star(Z)\big\}$$

In other words, we flip the predictions of $H^k_{S_m}$ if $k = 0$ and measure the misclassification loss afterwards. Note that this is a parametric loss; it has a bounded range and satisfies the Markov chain $S_m \to H \to L(\cdot;H)$. However, the expected generalization risk w.r.t. this parametric loss is $R_{gen}(\mathcal{L}) = \frac{1}{2}$ for all $m \geq 1$ because $R_{emp}(\mathcal{L}) = 0$ w.r.t. to this loss. Therefore, $\mathcal{L}$ does not generalize uniformly in expectation[2].

[2]We remark here that the Markov inequality cannot be used to provide a concentration bound for the learning algorithm in Proposition 1 even if the expected generalization risk goes to zero because the quantity $R_{emp}(H;S_m) - R_{true}(H)$ is not guaranteed to be non-negative. Indeed, this is precisely why the learning algorithm $\mathcal{L}$ constructed here generalizes in expectation but not in probability.

2 Proof of Theorem 2

We will first prove the inequality when $k = 2$. First, we write by definition:

$$\mathcal{J}(Z;\,(H_1,H_2)) = ||p(Z,H_1,H_2)\,,\;\; p(Z)\,p(H_1,H_2)||_{\mathcal{T}}$$

Using the fact that the total variation distance is related to the ℓ_1 distance by $||P\,,\;\; Q||_{\mathcal{T}} = \frac{1}{2}||P - Q||_1$, we have:

$$\begin{aligned}
\mathcal{J}(Z;\,(H_1,H_2)) &= \frac{1}{2}\Big|\Big|\; p(Z,H_1,H_2) - p(Z)\,p(H_1,H_2)\;\Big|\Big|_1 \\
&= \frac{1}{2}\Big|\Big|\; p(Z,H_1)\,p(H_2|Z,H_1) - p(Z)\,p(H_1)\,p(H_2|H_1)\;\Big|\Big|_1 \\
&= \frac{1}{2}\Big|\Big|\; \big[p(Z,H_1) - p(Z)\,p(H_1)\big]\cdot p(H_2|H_1) \\
&\qquad + p(Z,H_1)\;\cdot\big[p(H_2|Z,H_1) - p(H_2|H_1)\big]\;\Big|\Big|_1
\end{aligned}$$

Using the triangle inequality:

$$\begin{aligned}
&\mathcal{J}(Z_{trn};\,(H_1,H_2)) \\
&\quad\le \frac{1}{2}\Big|\Big|\big[p(Z,H_1) - p(Z)\,p(H_1)\big]\cdot p(H_2|H_1)\Big|\Big|_1 \\
&\qquad + \frac{1}{2}\;\Big|\Big|p(Z,H_1)\cdot\big[p(H_2|Z,H_1) - p(H_2|H_1)\big]\Big|\Big|_1
\end{aligned}$$

The above inequality is interpreted by expanding the ℓ_1 distance into a sum of absolute values of terms in the product space $\mathcal{Z}\times\mathcal{H}_1\times\mathcal{H}_2$, where $H_k \in \mathcal{H}_k$. Next, we bound each term on the right-hand side separately. For the first term, we note that:

$$\begin{aligned}
&\frac{1}{2}\Big|\Big|\;\big[p(Z,H_1) - p(Z)\,p(H_1)\big]\cdot p(H_2|H_1)\;\Big|\Big|_1 \\
&\qquad = \frac{1}{2}\Big|\Big|\;p(Z,H_1) - p(Z)\,p(H_1)\;\Big|\Big|_1 = \mathcal{J}(Z;\,H_1)
\end{aligned} \tag{1}$$

The equality holds by expanding the ℓ_1 distance and using the fact that $\sum_{H_2} p(H_2|H_1) = 1$.

However, the second term can be re-written as:

$$\begin{aligned}
&\frac{1}{2}\Big|\Big|\;p(Z,H_1)\;\cdot\big[p(H_2|Z,H_1) - p(H_2|H_1)\big]\;\Big|\Big|_1 \\
&\quad = \frac{1}{2}\Big|\Big|\;p(H_1)\cdot\big[p(H_2,Z|H_1) - p(Z|H_1)\,p(H_2|H_1)\big]\;\Big|\Big|_1 \\
&\quad = \mathbb{E}_{H_1}\big[||p(H_2,Z|H_1)\,,\;\; p(Z|H_1)\,p(H_2|H_1)||_{\mathcal{T}}\big] \\
&\quad = \mathcal{J}(Z;\,H_2\,|\,H_1)
\end{aligned} \tag{2}$$

Combining Eq. (1) and (2) yields the inequality:

$$\mathcal{J}(Z;\,(H_1,H_2)) \le \mathcal{J}(Z;\,H_1) + \mathcal{J}(Z;\,H_2\,|\,H_1) \tag{3}$$

Next, we use Eq. (3) to prove the general statement for all $k \ge 1$. By writing:

$$\begin{aligned}
\mathcal{J}(Z;\,(H_1,\ldots,H_k)) \le\; &\mathcal{J}(Z;\,H_k\,|\,(H_1,\ldots,H_{k-1})) \\
&+ \mathcal{J}(Z;\,(H_1,\ldots,H_{k-1}))
\end{aligned}$$

Repeating the same inequality on the last term on the right-hand side yields the statement of the theorem.

3 Proof of Proposition 2

We will use the following fact (Alabdulmohsin, 2015):

Fact 1 (Information Cannot Hurt)**.** *For any random variables X, Y, Z:*

$$\mathcal{J}(X;\,Y) \le \mathcal{J}(X;\,(Y,Z))$$

Now, by the triangle inequality:

$$\begin{aligned}
\mathcal{J}(A;\,C\,|\,B) &= \mathbb{E}_B||p(A|B)\cdot p(C|B)\,,\;\; p(A,C|B)||_{\mathcal{T}} \\
&= \mathbb{E}_{A,B}||p(C|B)\,,\;\; p(C|A,B)||_{\mathcal{T}} \\
&\le \mathbb{E}_{A,B}||p(C|B)\,,\;\; p(C)||_{\mathcal{T}} \\
&\qquad + \mathbb{E}_{A,B}||p(C)\,,\;\; p(C|A,B)||_{\mathcal{T}} \\
&= \mathbb{E}_B||p(C|B)\,,\;\; p(C)||_{\mathcal{T}} \\
&\qquad + \mathbb{E}_{A,B}||p(C)\,,\;\; p(C|A,B)||_{\mathcal{T}} \\
&= \mathcal{J}(B;\,C) + \mathcal{J}(C;\,(A,B))
\end{aligned}$$

Therefore:

$$\mathcal{J}(C;\,(A,B)) \ge \mathcal{J}(A;\,C\,|\,B) - \mathcal{J}(B;\,C)$$

Combining this with the following chain rule of Theorem 2:

$$\mathcal{J}(C;\,(A,B)) \le \mathcal{J}(A;\,C\,|\,B) + \mathcal{J}(B;\,C)$$

yields:

$$\Big|\mathcal{J}(C;\,(A,B)) - \mathcal{J}(A;\,C\,|\,B)\Big| \le \mathcal{J}(B;\,C)$$

Or equivalently:

$$\Big|\mathcal{J}(A;\,(B,C)) - \mathcal{J}(A;\,C\,|\,B)\Big| \le \mathcal{J}(A;\,B) \tag{4}$$

To prove the other inequality, we use Fact 1. We have:

$$\mathcal{J}(A;\,B) \le \mathcal{J}(A;\,(B,C)) \le \mathcal{J}(A;\,B) + \mathcal{J}(A;\,C\,|\,B),$$

where the first inequality follows from Fact 1 and the second inequality follows from the chain rule. Thus, we obtain the desired bound:

$$\Big|\mathcal{J}(A;\,(B,C)) - \mathcal{J}(A;\,B)\Big| \le \mathcal{J}(A;\,C\,|\,B) \tag{5}$$

Both Eq. 4 and Eq. 5 imply that the chain rule is tight. More precisely, the inequality can be made arbitrarily close to an equality when one of the two terms in the upper bound is chosen to be arbitrarily close to zero.

4 Proof of Theorem 3

We will use the following fact:

Fact 2. *Let $f : \mathcal{X} \to [0,1]$ be a function with a bounded range in the interval $[0,1]$. Let $p_1(x)$ and $p_2(x)$ be two different probability measures defined on the same space $\mathcal{X}$. Then:*

$$\left|\mathbb{E}_{X\sim p_1(x)} f(X) - \mathbb{E}_{X\sim p_2(x)} f(X)\right| \leq ||p_1(x)\,,\ p_2(x)||_{\mathcal{T}}$$

First, consider the following scenario. Suppose a learning algorithm $\mathcal{L}$ generates a hypothesis $H \in \mathcal{H}$ from some marginal distribution $p(h)$ *independently* of the sample S_m. Afterward, a sample $S_m \in \mathcal{Z}^m$ is observed, which comprises of m i.i.d. observations. Then, $\mathcal{L}$ selects $K \in \mathcal{K}$ according to $p(k|H, S_m)$.

In this scenario, we have:

$$\mathcal{J}(Z_{trn};\,(H,K)) = \mathcal{J}(Z_{trn};\,K\,|\,H),$$

where the equality follows from the chain rule in Theorem 2, the statement of Proposition 2, and the fact that $\mathcal{J}(Z_{trn};\,H) = 0$. The conditional variational information is written as:

$$\begin{aligned}&\mathcal{J}(Z_{trn};\,K\,|\,H)\\ &\quad= \mathbb{E}_H||p(Z_{trn})\cdot p(K|H)\,,\ p(Z_{trn},K|H)||_{\mathcal{T}},\end{aligned}$$

where we used the fact that $p(Z_{trn}|H) = p(Z_{trn})$. Next, by marginalization, the conditional distribution $p(K|H)$ is given by:

$$\begin{aligned}p(K|H) &= \mathbb{E}_{Z'_{trn}|H}[p(K|Z'_{trn},H)]\\ &= \mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)].\end{aligned}$$

where the expectation is taken with respect to the marginal distribution of observations $p(z)$. Similarly:

$$\begin{aligned}p(Z_{trn},K|H) &= p(Z_{trn}|H)\cdot p(K|Z_{trn},H)\\ &= p(Z_{trn})\cdot p(K|Z_{trn},H)\end{aligned}$$

Therefore:

$$\begin{aligned}&\mathcal{J}(Z_{trn};\,K\,|\,H)\\ &\quad= \mathbb{E}_H\mathbb{E}_{Z_{trn}}||\mathbb{E}_{Z'_{trn}}p(K|Z'_{trn},H)\,,\ p(K|Z_{trn},H)||_{\mathcal{T}}\end{aligned}$$

Next, for every value of H that is generated independently of the sample S_m, the variational information between $Z_{trn} \sim S_m$ and $K \in \mathcal{K}$ can be bounded using Theorem 3 in (Alabdulmohsin, 2015). This follows because H is selected independently of the sample S_m, and, hence, the i.i.d. property of the observations Z_i continue to hold. Therefore, we obtain:

$$\begin{aligned}&\mathbb{E}_H\mathbb{E}_{Z_{trn}}||\mathbb{E}_{Z'_{trn}}p(K|Z'_{trn},H)\,,\ p(K|Z_{trn},H)||_{\mathcal{T}}\\ &\qquad= \mathcal{J}(Z_{trn};\,K\,|\,H)\\ &\qquad\leq \sqrt{\frac{\log|\mathcal{K}|}{2m}}\end{aligned}\tag{6}$$

Because $p(K|Z_{trn},H)$ is arbitrary in our derivation, the above bound holds for any distribution of observations $p(z)$, any distribution $p(h)$, and any family of conditional distributions $p(k|Z_{trn},H)$.

Next, we return to the original setting where both $H \in \mathcal{H}$ and $K \in \mathcal{K}$ are chosen *according* to the sample S_m. We have:

$$\begin{aligned}&\mathcal{J}(Z_{trn};\,K\,|\,H)\\ &\quad= \mathbb{E}_H||p(Z_{trn}|H)\cdot p(K|H)\,,\ p(Z_{trn},K|H)||_{\mathcal{T}}\\ &\quad= \mathbb{E}_{H,Z_{trn}}||p(K|H)\,,\ p(K|Z_{trn},H)||_{\mathcal{T}}\\ &\quad= \mathbb{E}_{H,Z_{trn}}||\mathbb{E}_{Z'_{trn}|H}[p(K|Z'_{trn},H)]\,,\ p(K|Z_{trn},H)||_{\mathcal{T}}\\ &\quad\leq \mathbb{E}_{H,Z_{trn}}||\mathbb{E}_{Z'_{trn}|H}[p(K|Z'_{trn},H)]\,,\ \mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)]||_{\mathcal{T}}\\ &\qquad+ \mathbb{E}_{H,Z_{trn}}||\mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)]\,,\ p(K|Z_{trn},H)||_{\mathcal{T}}\end{aligned}\tag{7}$$

In the last line, we used the triangle inequality.

Next, we would like to bound the first term. Using the fact that the total variation distance is related to the ℓ_1 distance by $||P\,,\ Q||_{\mathcal{T}} = \frac{1}{2}||P-Q||_1$, we have:

$$\begin{aligned}&\mathbb{E}_{H,Z_{trn}}||\mathbb{E}_{Z'_{trn}|H}[p(K|Z'_{trn},H)]\,,\ \mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)]||_{\mathcal{T}}\\ &= \mathbb{E}_H\ ||\mathbb{E}_{Z'_{trn}|H}[p(K|Z'_{trn},H)]\,,\ \mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)]||_{\mathcal{T}}\\ &= \frac{1}{2}\,\mathbb{E}_H\sum_{K\in\mathcal{K}}\left|\mathbb{E}_{Z'_{trn}|H}[p(K|Z'_{trn},H)] - \mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)]\right|\\ &\leq \frac{1}{2}\,\mathbb{E}_H\sum_{K\in\mathcal{K}}||p(Z'_{trn}|H)\,,\ p(Z'_{trn})||_{\mathcal{T}}\\ &= \frac{1}{2}\sum_{K\in\mathcal{K}}\mathbb{E}_H||p(Z'_{trn}|H)\,,\ p(Z'_{trn})||_{\mathcal{T}}\\ &= \frac{1}{2}\sum_{K\in\mathcal{K}}\mathcal{J}(Z_{trn};\,H)\\ &= \frac{|\mathcal{K}|}{2}\,\mathcal{J}(Z_{trn};\,H)\end{aligned}\tag{8}$$

Here, the inequality follows from Fact 2.

Next, we bound the second term in Eq. 7. Using Fact 2 and our earlier result in Eq. 6:

$$\begin{aligned}&\mathbb{E}_{H,Z_{trn}}||\mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)]\,,\ p(K|Z_{trn},H)||_{\mathcal{T}}\\ &\qquad\leq \mathcal{J}(Z_{trn};\,H)\\ &\qquad\quad+ \mathbb{E}_H\mathbb{E}_{Z_{trn}}||\mathbb{E}_{Z'_{trn}}[p(K|Z'_{trn},H)]\,,\ p(K|Z_{trn},H)||_{\mathcal{T}}\\ &\qquad\leq \mathcal{J}(Z_{trn};\,H) + \sqrt{\frac{\log|\mathcal{K}|}{2m}}\end{aligned}\tag{9}$$

Combining all results in Eq. 7, 8, and 9:

$$\mathcal{J}(Z_{trn};\,K\,|\,H) \leq [1+\frac{|\mathcal{K}|}{2}]\,\mathcal{J}(Z_{trn};\,H) + \sqrt{\frac{\log|\mathcal{K}|}{2m}}\tag{10}$$

This along with the chain rule imply the statement of the theorem.

5 Proof of Theorem 4

Let $L(\cdot; H)$ be a parametric loss function and write:

$$\kappa(t) = p\Big\{\big|R_{emp}(H; S_m) - R_{true}(H)\big| \geq t\Big\} \qquad (11)$$

Consider the new pair of hypotheses (H, K), where:

$$K = \begin{cases} +1, & \text{if } R_{emp}(H; S_m) \geq R_{true}(H) + t \\ -1, & \text{if } R_{emp}(H; S_m) \leq R_{true}(H) - t \\ 0, & \text{otherwise} \end{cases}$$

Then, by Theorem 3, the uniform generalization risk in expectation for the composition of hypotheses (H, K) is bounded by $(7/2)\,\mathcal{J}(Z_{trn};\, H) + \sqrt{\frac{\log 3}{2m}}$. This holds uniformly across all parametric loss functions $L'(\cdot; H, K) \to [0, 1]$ that satisfy the Markov chain $S_m \to (H, K) \to L'(\cdot; H, K)$. Next, consider the parametric loss:

$$L'(Z; H, K) = \begin{cases} L(Z; H) & \text{if } K = +1 \\ 1 - L(Z; H) & \text{if } K = -1 \\ 0 & \text{otherwise} \end{cases}$$

Note that $L'(\cdot; H, K)$ is parametric with respect to the composition of hypotheses (H, K). Using Eq. 11, the generalization risk w.r.t $L'(\cdot; H, K)$ in expectation is, at least, as large as $t\,\kappa(t)$. Therefore, by Theorem 1 and Theorem 3, we have $t\,\kappa(t) \leq (7/2)\,\mathcal{J}(Z_{trn};\, H) + \sqrt{\frac{\log 3}{2m}}$. Because $\mathcal{J}(Z_{trn};\, H) \leq 1 - \mathbb{S}(\mathcal{L})$ by definition, the statement of the theorem immediately follows.

6 Proof of Proposition 3

Let $I(X; Y)$ denote the mutual information between X and Y and let $\mathbf{H}(X)$ denote the Shannon entropy of the random variable X measured in nats (i.e. using natural logarithms). We write $S_m = (Z_1, \ldots, Z_m)$. We have:

$$\begin{aligned} I(S_m; (H, K)) &= \mathbf{H}(S_m) - \mathbf{H}(S_m \mid H, K) \\ &= \sum_{i=1}^{m} \mathbf{H}(Z_i) - \sum_{i=1}^{m} \mathbf{H}(Z_i | H, K, Z_1, \ldots, Z_{i-1}) \\ &\geq \sum_{i=1}^{m} \mathbf{H}(Z_i) - \mathbf{H}(Z_i | H, K) \\ &= m I(Z_{trn}; H, K) \end{aligned}$$

The second line is the chain rule for entropy and the third lines follows from the fact that conditioning reduces entropy. We obtain:

$$I(Z_{trn}; H, K) \leq \frac{I(S_m; (H, K))}{m}$$

By Pinsker's inequality:

$$\mathcal{J}(Z_{trn};\, (H, K)) \leq \sqrt{\frac{I(Z_{trn}; (H, K))}{2}} \leq \sqrt{\frac{I(S_m; (H, K))}{2m}}$$

Using the chain rule for mutual information:

$$\begin{aligned} \mathcal{J}(Z_{trn};\, (H, K)) &\leq \sqrt{\frac{I(S_m; (H, K))}{2m}} \\ &= \sqrt{\frac{I(S_m; H) + I(S_m; K | H)}{2m}} \\ &\leq \sqrt{\frac{I(S_m; H) + \mathbf{H}(K)}{2m}} \\ &\leq \sqrt{\frac{I(S_m; H) + \log |\mathcal{K}|}{2m}} \end{aligned}$$

The desired bound follows by applying the same proof technique of Theorem 4 on the last uniform generalization bound, and using the fact that $\log 3 < 2$.

7 Proof of Corollary 1

First, we note that for any two adjacent samples S and S', we have:

$$p(H|S) - p(H|S') \leq (e^{\epsilon} - 1)\; p(H|S') + \delta$$

This follows by definition of differential privacy. Similarly, we have:

$$\begin{aligned} p(H|S) - p(H|S') &\geq (e^{-\epsilon} - 1)\; p(H|S') - e^{-\epsilon}\delta \\ &= -\Big[(1 - e^{-\epsilon})\; p(H|S') + e^{-\epsilon}\delta\Big] \\ &\geq -e^{\epsilon}\; \Big[(1 - e^{-\epsilon})\; p(H|S') + e^{-\epsilon}\delta\Big] \\ &= -\Big[(e^{\epsilon} - 1) p(H|S') + \delta\Big] \end{aligned}$$

Both results imply that:

$$\big|p(H|S) - p(H|S')\big| \leq (e^{\epsilon} - 1) p(H|S') + \delta \qquad (12)$$

Writing:

$$\begin{aligned} \mathcal{J}(Z_{trn};\, H) &= ||p(Z_{trn}, H)\,,\;\; p(Z_{trn}) \cdot p(H)||_{\mathcal{T}} \\ &= \mathbb{E}_{Z_{trn}}\; ||p(H|Z_{trn})\,,\;\; p(H)||_{\mathcal{T}} \\ &= \frac{1}{2}\mathbb{E}_{Z_{trn}}\Big|\Big|\mathbb{E}_{Z'_{trn}}\big[p(H|Z_{trn}) - p(H|Z'_{trn})\big]\Big|\Big|_1 \\ &\leq \frac{1}{2}\mathbb{E}_{Z_{trn}, Z'_{trn}}\Big|\Big|p(H|Z_{trn}) - p(H|Z'_{trn})\Big|\Big|_1 \end{aligned}$$

The last inequality follows by convexity. Next, let S_{m-1} be a sample that contains $m - 1$ observations

drawing i.i.d. from $p(z)$. Then:

$$\begin{aligned}\mathcal{J}(Z_{trn};\, H) &\leq \frac{1}{2}\mathbb{E}_{Z_{trn},Z'_{trn}}\Big|\Big|p(H|Z_{trn}) - p(H|Z'_{trn})\Big|\Big|_1 \\ &= \frac{1}{2}\mathbb{E}_{Z_{trn},Z'_{trn}}\Big|\Big|\mathbb{E}_{S_{m-1}}\big[p(H|Z_{trn}, S_{m-1}) \\ &\qquad - p(H|Z'_{trn}, S_{m-1})\big]\Big|\Big|_1 \\ &\leq \frac{1}{2}\mathbb{E}_{S,S'}\Big|\Big|p(H|S) - p(H|S')\Big|\Big|_1,\end{aligned}$$

where S, S' are two adjacent samples.

Next, we expand the ℓ_1 distance and use Eq 12:

$$\begin{aligned}\mathcal{J}(Z_{trn};\, H) &\leq \frac{1}{2}\mathbb{E}_{S,S'}\Big|\Big|p(H|S) - p(H|S')\Big|\Big|_1 \\ &= \frac{1}{2}\mathbb{E}_{S,S'}\sum_{H\in\mathcal{H}}\big|p(H|S) - p(H|S')\big| \\ &\leq \frac{1}{2}\mathbb{E}_{S,S'}\sum_{H\in\mathcal{H}}\big[(e^{\epsilon}-1)p(H|S') + \delta\big] \\ &= \frac{e^{\epsilon}-1+\delta}{2}\end{aligned}$$

Finally, the desired bound follows by combining the last inequality with Theorem 4.

8 Proof of Corollary 2

It has been shown in Alabdulmohsin (2015) that the supremum generalization risk is achieved (arbitrarily well) using the following binary-valued loss:

$$L^{\star}(z; H) = \mathbb{I}\big\{p(Z_{trn} = z|H) \geq p(Z_{trn} = z)\big\} \quad (13)$$

Therefore, if an algorithm is (ϵ, δ) robustly generalizing, let the adversary $\mathcal{A}$ (or equivalently the parametric loss $L(\cdot; H)$) be fixed to the one given by Eq. 13. Hence, we have by definition of robust generalization:

$$p\Big\{\big|\mathbb{E}_{Z\sim p(z)}L^{\star}(Z; H) - \frac{1}{m}\sum_{Z_i\in S}L^{\star}(Z_i; H)\big| \leq \epsilon\Big\} \geq 1-\delta, \quad (14)$$

Therefore:

$$\Big|\mathbb{E}_{S,H}\big[\mathbb{E}_{Z\sim p(z)}L^{\star}(Z; H) - \frac{1}{m}\sum_{Z_i\in S}L^{\star}(Z_i; H)\big]\Big| \leq \epsilon + \delta$$

Because $L^{\star}(\cdot; H)$ achieves the maximum possible generalization risk in expectation (Alabdulmohsin, 2015), we have the uniform generalization bound $\mathcal{J}(Z_{trn};\, H) \leq \epsilon+\delta$. Hence, (ϵ, δ) robust generalization implies a uniform generalization at the rate $\epsilon + \delta$.

The proof of the converse follows from our concentration bound in Theorem 4, which shows that uniform generalization in expectation implies a generalization in probability. In particular, any algorithm that generalizes uniformly with rate τ is (ϵ, γ) robustly generalizing, with $\gamma = (7/2)(\tau + \sqrt{\log 3/(49m)})/\epsilon$.

9 Proof of Theorem 5

Before we prove the statement of the theorem, we begin with the following lemma:

Lemma 1. *Let the observation space $\mathcal{Z}$ be the interval $[0, 1]$, where $p(z)$ is continuous in $[0, 1]$. Let $H \subseteq S_m : |H| = k$ be a set of k examples picked at random without replacement from the sample S_m. Then $\mathcal{J}(Z_{trn};\, H) = \frac{k}{m}$.*

Proof. First, we note that $p(Z_{trn}|H)$ is a *mixture* of two distributions: one that is uniform in H with probability k/m, and the original distribution $p(z)$ with probability $1 - k/m$. By Jensen's inequality, we have $\mathcal{J}(Z_{trn};\, H) \leq k/m$. Second, let the parametric loss be $L(\cdot; H) = \mathbb{I}\{Z \in H\}$. Then, $|R_{gen}(\mathcal{L})| = \frac{k}{m}$. By Theorem 1, we have $\mathcal{J}(Z_{trn};\, H) \geq |R_{gen}(\mathcal{L})| = k/m$. Both bounds imply the statement of the lemma. □

Now, we prove Theorem 5. Consider the example where $\mathcal{Z} = [0, 1]$ and suppose that the observations $Z \in \mathcal{Z}$ have a continuous marginal distribution. Because t is a rational number, let the sample size m be chosen such that $k = t\, m$ is an integer.

Let $\{Z_1, \ldots, Z_m\}$ be the training set, and let the hypothesis H be given by $H = \{Z_1, \ldots, Z_k\}$ with some probability $\delta > 0$ and $H = \{\}$ otherwise. Here, the k instances $Z_i \in H$ are picked uniformly at random without replacement from the sample S_m. To determine the variational information between Z_{trn} and H, we consider the two cases:

1. If $H \neq \{\}$, then $||p(Z_{trn})\,,\ p(Z_{trn}|H)||_{\mathcal{T}} = t$ as proved in Lemma 1. This happens with probability δ by construction.
2. If $H = \{\}$ then $p(Z_{trn}|H) = p(Z_{trn})$. Hence, we have $||p(Z_{trn})\,,\ p(Z_{trn}|H = \{\})||_{\mathcal{T}} = 0$. This happens with probability $1 - \delta$.

So, by combining the two cases above, we deduce that:

$$\mathcal{J}(Z_{trn};\, H) = \mathbb{E}_H\, ||p(Z_{trn})\,,\ p(Z_{trn} \mid H)||_{\mathcal{T}} = t\ \delta.$$

Therefore, $\mathcal{L}$ generalizes uniformly with the rate $t\delta$. Next, let the loss $L(\cdot; H)$ be given by $L(Z\ ;\ H) = \mathbb{I}\big\{Z \in H\big\}$. With this loss:

$$\begin{aligned}p\Big\{\big|R_{emp}(H; S_m) - R_{true}(H)\big| = t\Big\} &= \delta \\ &= \frac{\mathcal{J}(Z_{trn};\, H)}{t},\end{aligned}$$

which is the statement of the theorem.

10 Proof of Proposition 4

This proposition is proved in Corollary 1.

References

Alabdulmohsin, I. (2015). Algorithmic stability and uniform generalization. In *NIPS*, pages 19–27.