A summary of the paper entitled:

"Chaining Mutual Information and Tightening Generalization Bounds"

By:
Amir R. Asadi

The paper [1], titled above, gives a technique for deriving multi-scale and algorithm-dependent generalization bounds for learning algorithms by combining ideas from high dimensional probability and from information theory. Various generalization bounds have been proposed in statistical learning theory, such as the basic union bound over the hypothesis set, the refined union bound, Rademacher complexity, chaining and VC-dimension [2, 3]; and algorithm-dependent bounds such as PAC-Bayesian bounds [4], uniform stability [5], compression bounds [6], and recently, the mutual information bound [7]. Two pitfalls among the key limitations of current bounds are the following:

A. Ignoring the dependencies between the hypotheses. Generalization bounds which are not based on any metric on the hypothesis set may not exploit the dependencies between the hypotheses. On the other hand, chaining, originated from the work of Kolmogorov and Dudley, is a powerful technique in high dimensional probability for bounding the expected suprema of random processes while taking into account the dependencies between their random variables in a multi-scale manner. It is the method for proving the tightest generalization bound using VC-dimension [8, 9]. The basic idea of chaining is to first describe the dependencies between the random variables of a random process $\{X_t\}_{t\in T}$ by a metric d on the set T, then to discretize T and to approximate the supremum of the random process by approximating the maxima over successively refined finite discretizations, using union bounds at each step, and by using the notions of *ϵ-nets* and *covering numbers* [10]. Here, we state Dudley's inequality, a fundamental result based on the chaining technique. For a metric space (T, d), let $N(T, d, \epsilon)$ denote the covering number of (T, d) at scale ϵ. Dudley [11] showed that when $\{X_t\}_{t\in T}$ is a separable subgaussian process on the bounded metric space (T, d), then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T, d, 2^{-k})}. \tag{1}$$

Notice that Dudley's inequality is a multi-scale bound. However, it is bounding $\mathbb{E}\left[\sup_{t\in T} X_t\right]$, hence is a uniform bound and does not take into account the algorithm.

B. Ignoring the dependence between the algorithm input (data) and output. Generalization bounds based on Rademacher complexity and VC-dimension only depend on the hypothesis set and not on the algorithm, effectively rendering them too pessimistic and vacuous for practical algorithms. If $\mathcal{H} = \{h_t : t \in T\}$ denotes the hypothesis set and for every $t \in T$, X_t denotes the generalization error of hypothesis h_t and W denotes the index of the chosen hypothesis by the algorithm, then to upper bound the expected generalization error $\mathbb{E}[X_W]$, one uses

$$\mathbb{E}[X_W] \leq \mathbb{E}\left[\sup_{t\in T} X_t\right], \tag{2}$$

and aims at upper bounding $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ with these bounds, hence giving a uniform bound over the generalization errors of the entire hypothesis set. However, the generalization error

X_W of the specific hypothesis W selected by the algorithm may be much smaller than the $\sup_{t\in T} X_t$. In other words, these bounds are not taking into account the input-output relation of the learning algorithm, and uniform bounding seems to be too stringent for these applications. A recent method to tighten (2) is given with the following algorithm-dependent bound which uses the mutual information between the output index and the random process (see [7, 12]): Suppose that $\{X_t\}_{t\in T}$ is a random process and T an arbitrary set. Assume that X_t is σ^2-subgaussian and $\mathbb{E}[X_t] = 0$ for every $t \in T$, and let W be a random variable taking values on T. Then

$$|\mathbb{E}[X_W]| \leq \sqrt{2\sigma^2 I(W;\{X_t\}_{t\in T})}. \tag{3}$$

Inequality (3) implies that whenever the mutual information between the output index and the random process is small, then the generalization error/bias is small too. However, it is not difficult to see that this condition is not necessary, as the bound does not exploit the dependencies between the random variables of the process (it is not based on any metric on T). That is, an algorithm may have high mutual information between its input and output, but still generalize well. A simple example which makes (3) extremely loose is the following: Consider the Gaussian process $X_t \triangleq \langle t, G^2\rangle$, $t \in T$ where $G^2 = (G_1, G_2)$ has independent standard normal components and $T \triangleq \{t \in \mathbb{R}^2 : |t|_2 = 1\}$. The process $\{X_t\}_{t\in T}$ can also be expressed according to the phase of each point $t \in T$, i.e. the unique number $\phi \in [0, 2\pi)$ such that $t = (\sin\phi, \cos\phi)$. Assume that the indices are in the phase form and let $W \triangleq \left(\text{argmax}_{\phi\in[0,2\pi)} X_\phi\right) \oplus Z \pmod{2\pi}$, where the noise Z is independent from $\{X_t\}_{t\in T}$, and has an atom with probability mass ϵ on 0, and has $1-\epsilon$ probability uniformly distributed on $(-\pi, \pi)$. We have $\mathbb{E}[X_W] = \epsilon\sqrt{\frac{\pi}{2}}$, which can be made arbitrarily close to zero by choosing ϵ arbitrarily small, however, $I(W;\{X_t\}_{t\in T}) = \infty$ for all $\epsilon > 0$.

Combining chaining with mutual information In [1], the ideas of the chaining technique and the mutual information bound are combined to obtain a technique for bounding the expected generalization error which takes into account the dependencies between the hypotheses in a multi-scale manner, as well as the dependence between output and input of the algorithm. The technique of "chaining mutual information" can be interpreted as an algorithm-dependent version of chaining, extending a result of [13] by taking into account such dependencies. Before giving the main results, we give the definition of "increasing sequence of ϵ-partitions". We call a partition $\mathcal{P} = \{A_1, A_2, ..., A_m\}$ of the set T an *ϵ-partition* of the metric space (T, d) if for all $i = 1, 2, ..., m$, A_i can be contained within a ball of radius ϵ. A sequence of partitions $\{\mathcal{P}_k\}_{k=m}^{\infty}$ of a set T is called an *increasing sequence* if for all $k \geq m$ and each $A \in \mathcal{P}_{k+1}$, there exists $B \in \mathcal{P}_k$ such that $A \subseteq B$. For any such sequence and any $t \in T$, let $[t]_k$ denote the unique set $A \in \mathcal{P}_k$ such that $t \in A$.

Theorem 1. *Assume that $\{X_t\}_{t\in T}$ is a separable subgaussian process on the bounded metric space (T, d). Let $\{\mathcal{P}_k\}_{k=k_1(T)}^{\infty}$ be an increasing sequence of partitions of T, where $k_1(T)$ is an integer such that $2^{-(k_1(T)-1)} \geq \text{diam}(T)$ and for each $k \geq k_1(T)$, $\mathcal{P}_k$ is a 2^{-k}-partition of (T, d). Then*

$$\mathbb{E}[X_W] \leq 3\sqrt{2} \sum_{k=k_1(T)}^{\infty} 2^{-k} \sqrt{I([W]_k;\{X_t\}_{t\in T})}. \tag{4}$$

Let $\mathcal{X}$ and $\mathcal{Y}$ be the instances/labels domain and $\mathcal{H} = \{h_w : w \in \mathcal{W}\}$ be the hypothesis set where the hypotheses are indexed by an index set $\mathcal{W}$. Let $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}^+$ be a loss

function. A learning algorithm receives the training set $S = (Z_1, Z_2, ..., Z_n)$ of n examples with i.i.d. random elements drawn from $\mathcal{X} \times \mathcal{Y}$ with distribution μ, and picks an element $h_W \in \mathcal{H}$ as the output hypothesis according to a random transformation $P_{W|S}$ (thus, we are allowing randomized algorithms). For any $w \in \mathcal{W}$, let $L_\mu(w) \triangleq \mathbb{E}[\ell(h_w, Z)]$ denote the statistical risk of hypothesis h_w, where $Z \sim \mu$. For a given training set S, the empirical risk of hypothesis h_w is defined as $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(h_w, Z_i)$, and the generalization error of hypothesis h_w (dependent on the training set) is defined as $\mathrm{gen}(w) \triangleq L_\mu(w) - L_S(w)$. Averaging with respect to the joint distribution $P_{S,W} = \mu^{\otimes n} P_{W|S}$, we denote the expected generalization error by $\mathrm{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}[L_\mu(W) - L_S(W)]$. The following result follows from Theorem 1, using the data processing inequality:

Theorem 2. *Assume that $\{\mathrm{gen}(w)\}_{w\in\mathcal{W}}$ is a separable subgaussian process on the bounded metric space $(\mathcal{W}, d)$. Let $\{\mathcal{P}_k\}_{k=k_1(\mathcal{W})}^{\infty}$ be an increasing sequence of partitions of $\mathcal{W}$, where $k_1(\mathcal{W})$ is an integer such that $2^{-(k_1(\mathcal{W})-1)} \geq \mathrm{diam}(\mathcal{W})$ and for each $k \geq k_1(\mathcal{W})$, $\mathcal{P}_k$ is a 2^{-k}-partition of $(\mathcal{W}, d)$. Then*

$$\mathrm{gen}(\mu, P_{W|S}) \leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I([W]_k; S)}. \tag{5}$$

References

[1] Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, pages 7234–7243, 2018.

[2] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.

[3] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[4] David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[5] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

[6] Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. Technical report, University of California, Santa Cruz, 1986.

[7] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240, 2016.

[8] Ramon van Handel. Probability in high dimension. *[Online]. Available:* `https://www.princeton.edu/~rvan/APC550.pdf`, Dec. 21 2016.

[9] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[11] Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[12] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

[13] Xavier Fernique. Evaluations de processus Gaussiens composes. In *Probability in Banach Spaces*, pages 67–83. Springer, 1976.