A Summary of Some Interesting Bounds in Estimation and Learning

Qiang Ning

***Abstract*—These review notes serve as a guide for myself to some bounds of interest in the estimation theory and learning theory, including Cramér-Rao Bound (CRB), concentration inequalities, Vapnic-Chervonenkis (VC) theory, probably approximately correct (PAC) learning, and the Johnson-Lindenstrauss (JL) lemma.**

***Index Terms*—Cramér-Rao Bounds, Concentration Inequalities, VC Dimension, PAC Learning**

I. Introduction

THIS review was generated in February, 2015. When we were preparing an extended version of our ISBI paper "Spectral Estimation for Magnetic Resonance Spectroscopic Imaging with Spatial Sparsity Constraints", how to characterize the performance of our method aroused our interest. Therefore, we spent some extra time digging into some interesting bounds, especially in the field of estimation and learning. This review is just a summary of these bounds.

II. Bounds for Parameter Estimation

A. Basic Settings

Let the parameters of an estimator be a k dimensional vector $\mathbf{x}$, and measurement data be an n dimensional random vector $\mathbf{Y}$, which follows the probability distribution function $f(\mathbf{Y}|\mathbf{x})$ or $f_{\mathbf{x}}(\mathbf{y})$. $\hat{\mathbf{X}}(\mathbf{Y})$ denotes the estimator.

B. Cramér-Rao Bounds

The Cramer-Rao Bound (CRB) that we usually used for unbiased estimators is

$$\mathbf{C}_E(\mathbf{x}) \geq \mathbf{J}^{-1}(\mathbf{x}), \tag{1}$$

where

$$\mathbf{C}_E(\mathbf{x}) = E[(\mathbf{x}-\hat{\mathbf{X}}(\mathbf{Y}))(\mathbf{x}-\hat{\mathbf{X}}(\mathbf{Y}))^T]$$

is the error correlation matrix of X, and

$$\mathbf{J}(\mathbf{x}) = E_{\mathbf{Y}}[\bigtriangledown_{\mathbf{x}} \ln f(\mathbf{Y}|\mathbf{x})(\bigtriangledown_{\mathbf{x}} \ln f(\mathbf{Y}|\mathbf{x}))^T], \tag{2}$$

or equivalently (which is proved in [1]),

$$\mathbf{J}(\mathbf{x}) = -E_{\mathbf{Y}}[\bigtriangledown_{\mathbf{x}} \bigtriangledown_{\mathbf{x}}^T \ln f(\mathbf{Y}|\mathbf{x})], \tag{3}$$

is the so-called Fisher Information in $\mathbf{Y}$ about the parameters in $\mathbf{X}$. Note in (2) and (3), $f(\mathbf{Y}|\mathbf{x})$ is the conditional probability distribution function of measurement $\mathbf{Y}$ on parameters $\mathbf{x}$, and the gradient operator is defined as

$$\bigtriangledown_x = \left[\begin{array}{cccc} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \cdots & \frac{\partial}{\partial x_k} \end{array} \right]^T.$$

In practice, the Fisher Information can be calculated entry-wisely in the following two ways.

$$\begin{aligned} J_{i,j}(\mathbf{x}) &= E_{\mathbf{Y}}[\frac{\partial}{\partial x_i} \ln f(\mathbf{Y}|\mathbf{x}) \frac{\partial}{\partial x_j} \ln f(\mathbf{Y}|\mathbf{x})] \\ &= -E_{\mathbf{Y}}[\frac{\partial^2}{\partial x_i \partial x_j} \ln f(\mathbf{Y}|\mathbf{x})]. \end{aligned}$$

Given this bound, we have the estimator variance of each parameter,

$$E[(x_i - \hat{x}_i(\mathbf{Y}))^2] \geq [J^{-1}]_{ii}(\mathbf{x}),$$

which is useful in the analysis of the efficiency of an estimator.

A more general form is

Theorem 1 (Cramér-Rao Bound)**.** *The following inequality holds true:*

$$\mathbf{C}_E(\mathbf{x}) \geq \mathbf{b}(\mathbf{x})\mathbf{b}^T(\mathbf{x}) + (\mathbf{I}_k - \bigtriangledown_x^T \mathbf{b}(\mathbf{x}))\mathbf{J}^{-1}(\mathbf{x})(\mathbf{I}_k - \bigtriangledown_x^T \mathbf{b}(\mathbf{x}))^T, \tag{4}$$

where $\mathbf{b}(\mathbf{x})$ *is the bias of this estimator,*

$$\mathbf{b}(\mathbf{x}) = \mathbf{x} - E(\hat{\mathbf{X}}(\mathbf{Y})).$$

It is obvious that when the estimator is unbiased, that is, $\mathbf{b}(\mathbf{x}) = 0$*, (4) degenerates to (1).*

C. Barankin Bounds

The Barankin bounds considers the variance of an estimator of $g(\mathbf{x})$, which is a real-valued function of $\mathbf{x}$. Let $\hat{g}(\mathbf{x})$ be an unbiased estimator of $g(\mathbf{x})$. Then the variance of $\hat{g}(\mathbf{x})$ is bounded from the below.

Lemma 1 (Barankin Bound)**.** *The following inequality holds true for any finite* $\mathbf{x}^{(i)}, a_i, i = 1, \ldots, p$*,*

$$\sigma^2(\hat{g}) \geq \frac{\{\sum_{i=1}^p a_i [g(\mathbf{x}^{(i)}) - g(\mathbf{x})]\}^2}{\int [\sum_{i=1}^p a_i L(\mathbf{x}^{(i)}, \mathbf{x})]^2 f(\mathbf{y}|\mathbf{x}) d\mathbf{y}}, \tag{5}$$

where $L(\mathbf{x}^{(i)}, \mathbf{x}) = \frac{f(\mathbf{y}|\mathbf{x}^{(i)})}{f(\mathbf{y}|\mathbf{x})}$.

Barankin has shown that if an unbiased estimator of $\mathbf{x}$ exists, then there exists an unbiased estimator that achieves the Barankin bound above. This estimator is usually dependent on the specific value of $\mathbf{x}$, hence the name "locally best unbiased estimates" [3].

Based on this basic result, McAulay et al. have shown the following theorem [3].

Q. Ning is with the Department of Electrical and Computer Engineering and the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL. Email: qning2@illinois.edu.

Theorem 2. *$\hat{\mathbf{X}}$ denotes any unbiased estimator of $\mathbf{x}$. The following inequality holds true.*

$$\mathbf{\Sigma}(\hat{\mathbf{X}}) \geq \mathbf{\Lambda}^{-1} + (\mathbf{\Phi} - \mathbf{\Lambda}^{-1}\mathbf{A})\mathbf{\Delta}^{-1}(\mathbf{\Phi} - \mathbf{\Lambda}^{-1}\mathbf{A})^T,$$

where $\mathbf{\Delta} = \mathbf{B} - \mathbf{A}^T\mathbf{\Lambda}^{-1}\mathbf{A}$, $\mathbf{\Phi} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}]$,

$$\Lambda_{ij} = \int \frac{\partial \ln f(\mathbf{y}|\mathbf{x})}{\partial x_i} \frac{\partial \ln f(\mathbf{y}|\mathbf{x})}{\partial x_j} f(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \quad i,j = 1, \ldots, k,$$

$$A_{ij} = \int \frac{\partial \ln f(\mathbf{y}|\mathbf{x})}{\partial x_i} L(\mathbf{x}^{(j)}, \mathbf{x}) f(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \quad i = 1, \ldots, k, j = 1, \ldots, p,$$

$$B_{ij} = \int L(\mathbf{x}^{(i)}, \mathbf{x}) L(\mathbf{x}^{(j)}, \mathbf{x}) f(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \quad i,j = 1, \ldots, p.$$

To clarify, $\{\mathbf{x}^{(i)}\}_{i=1}^p$ is the set of test points, and x_i is the i-th element of vector $\mathbf{x}$. When there is no test points ($p = 0$), the Barankin bound turns out to be the CRB for unbiased esimation. When $p > 0$, the Barankin bounds are in general an improvement on the CRB, since $\mathbf{\Delta}$ is positive definite and $(\mathbf{\Phi} - \mathbf{\Lambda}^{-1}\mathbf{A})\mathbf{\Delta}^{-1}(\mathbf{\Phi} - \mathbf{\Lambda}^{-1}\mathbf{A})^T$ is positive semidefinite.

D. Chapman-Robbins Bounds

Chapman and Robbins derived another bound on the variance of an estimated parameter for any unbiased estimator [2]. They followed the basic steps of the proof of CRB, but avoided the need to differentiate under integral signs or to perform Schur decomposition. There results might also be obtained from Barankin's bounds [2][3][4]. In [2], only the case of unbiased estimation of a single real parameter was given as follows.

Lemma 2. *Given a parameter $x \in \mathbb{R}$, assume the measurement data $Y \in \mathbb{R}^n$ have a probability distribution $f_x(y)$. Then for any unbiased estimator $\hat{X}(Y)$, the following inequality holds true.*

$$Var(\hat{X}) \geq \frac{1}{E[J]}, \tag{6}$$

where $J = \frac{1}{h^2}\{[\frac{f_{x+h}(y)}{f_x(y)}]^2 - 1\}$. Since (6) holds true for all $h \neq 0$, we actually can reach a tighter bound by achieving the infimum of $E[J]$:

$$Var(\hat{X}) \geq \frac{1}{\inf_{h \neq 0} E[J]}.$$

The lemma can be obtained by choosing two test points in (5): $a_1 = \frac{1}{h}, a_2 = -\frac{1}{h}, x^{(1)} = x + h, x^{(2)} = x, h \neq 0$. The lemma can also be extended to the multi-parameter case.

Theorem 3 (Chapman-Robbins Bound)**.** *Given a parameter $\mathbf{x} \in \mathbb{R}^k$, assume the measurement data $\mathbf{Y} \in \mathbb{R}^n$ have a probability distribution $f_{\mathbf{x}}(\mathbf{y})$. Then for any estimator $\hat{\mathbf{X}}(\mathbf{Y})$, the following inequality holds true.*

$$\mathbf{\Sigma}(\hat{\mathbf{X}}) \geq [\delta\mathbf{m}_{\mathbf{x}}]^T \left(E\left[\left[\frac{\delta f_{\mathbf{x}}}{f_{\mathbf{x}}}\right] \left[\frac{\delta f_{\mathbf{x}}}{f_{\mathbf{x}}}\right]^T \right] \right)^{\dagger} [\delta\mathbf{m}_{\mathbf{x}}],$$

where $\mathbf{\Sigma}(\hat{\mathbf{X}})$ is the covariance matrix of $\hat{\mathbf{X}}(\mathbf{Y})$, $\mathbf{m}_{\mathbf{x}} \equiv E[\hat{\mathbf{X}}]$, "$\dagger$" denotes the Moore-Penrose pseudoinverse,

$$\delta\mathbf{m}_{\mathbf{x}} \equiv \left[\frac{\mathbf{m}_{\mathbf{x}+\mathbf{v}_1} - \mathbf{m}_{\mathbf{x}}}{\|\mathbf{v}_1\|}, \ldots, \frac{\mathbf{m}_{\mathbf{x}+\mathbf{v}_p} - \mathbf{m}_{\mathbf{x}}}{\|\mathbf{v}_p\|} \right]^T,$$

and

$$\delta f_{\mathbf{x}}(\mathbf{y}) \equiv \left[\frac{f_{\mathbf{x}+\mathbf{v}_1}(\mathbf{y}) - f_{\mathbf{x}}(\mathbf{y})}{\|\mathbf{v}_1\|}, \ldots, \frac{f_{\mathbf{x}+\mathbf{v}_p}(\mathbf{y}) - f_{\mathbf{x}}(\mathbf{y})}{\|\mathbf{v}_p\|} \right]^T,$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^k$ are p arbitrary vectors. Again this bound can be improved by achieving the supremum of the right hand side over all possible $\mathbf{v}_1, \ldots, \mathbf{v}_p$.

It can be seen that the Chapman-Robbins bound (by taking the supremum over all possible deviations) is at least as sharp as the CRB (by taking the derivatives) [2], but it is also more difficult to compute in general.

E. Constrained Cramér-Rao Bounds

There are many cases in practice that we have constraints when estimating parameters. Intuitively, imposing constraints can help reduce the variance of an estimator. However, the conventional CRB does not take the constraints into account. As a result, the corresponding CRB is too "pessimistic", and the achievable estimator variance can even be lower (better) than the CRB. Given this situation, we can directly derive the so-called constrained CRB using the Chapman-Robbins bound [5].

Theorem 4 (Constrained CRB)**.** *Let the parameter $\mathbf{x}$ to be estimated lie in a constrained space $\mathcal{X} \subset \mathbb{R}^k$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_p$ be p linear independent vectors with sufficient small lengths such that $\mathbf{x} + \mathbf{v}_i \in \mathcal{X}, i = 1, \ldots, p$. Assume the measurement data $\mathbf{Y} \in \mathbb{R}^n$ have a probability distribution $f_{\mathbf{x}}(\mathbf{y})$. Then for any estimator $\hat{\mathbf{X}}(\mathbf{Y})$ with mean $\mathbf{m}_{\mathbf{x}}$, the following inequality holds true:*

$$\mathbf{\Sigma}(\hat{\mathbf{X}}) \geq \limsup_{\substack{\mathbf{x}+\mathbf{v}_i \in \mathcal{X}, \|\mathbf{v}_i\| \to 0 \\ i=1,\ldots,p}} B_c,$$

where B_c is the defined as the Chapman-Robbins bound over $\mathbf{v}_1, \ldots, \mathbf{v}_p$. If some additional conditions hold (see [5]), then

$$\mathbf{\Sigma}(\hat{\mathbf{X}}) \geq [\nabla_x \mathbf{m}_{\mathbf{x}}]^T \mathbf{A} [\mathbf{A}^T \mathbf{J} \mathbf{A}]^{\dagger} \mathbf{A}^T [\nabla_x \mathbf{m}_{\mathbf{x}}],$$

where $\mathbf{A}$ is any matrix whose column space is the same as span$\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$, and $\mathbf{J}$ is the Fisher information matrix.

Based on the constrained CRB, one can characterize the performance of an estimator given sparsity constraint [6] or low-rank constraint [7], etc.

III. Concentration of Measure Inequalities

A. Basic Inequalities

Theorem 5 (Markov Inequality)**.** *For any nonnegative random variable X and $t > 0$, we have*

$$P\{X \geq t\} \leq \frac{E[X]}{t}.$$

Theorem 6 (Chebyshev's Inequality)**.** *For any random variable X and $t > 0$, we have*

$$P\{|X - E[X]| \geq t\} \leq \frac{E[|X - E[X]|^2]}{t^2}.$$

Theorem 7 (Chernoff's Bound)**.** *Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables with finite mean μ. Let $S_n = X_1 + X_2 + \cdots + X_n$. For $a > \mu$, we have*

$$P\{\frac{S_n}{n} \geq a\} \leq exp(-n[\theta a - \ln M(\theta)]),$$

where $M(\theta) = E[e^{\theta X_1}], \theta > 0$.

Usually since the left hand side of the Chernoff's bound does not dependent on θ, we can minimize the right hand side over θ to achieve tighter bounds. Note here the condition $a > \mu$ is required. If $P\{\frac{S_n}{n} \leq a\}, a < \mu$ is desired, we can construct $Y_n = -X_n, n = 1, 2, \dots$, and then apply this bound.

B. Other Inequalities

All of these inequalities below come from Lugosi [8].

Theorem 8 (Chebyshev-Cantelli Inequality)**.** *Let $t \geq 0$. Then*

$$P\{X - E[X] \geq t\} \leq \frac{Var(X)}{Var(X) + t^2}.$$

Theorem 9 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables with finite mean m. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then $S_n/n \xrightarrow{p} m$.*

Theorem 10 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables with finite mean m. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then $S_n/n \xrightarrow{a.s.} m$.*

Lemma 3 (Hoeffding's Inequality)**.** *Let X be a random variable with $E[X] = 0, a \leq X \leq b$. Then for $s > 0$,*

$$E[e^{sX}] \leq e^{s^2(b-a)^2/8}.$$

Theorem 11 (Hoeffding's Tail Inequality)**.** *Let $X_1, X_2, \dots, X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ with probability one. Then for any $t > 0$ we have*

$$P\{S_n - E[S_n] \geq t\} \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2},$$
$$P\{S_n - E[S_n] \leq -t\} \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

Theorem 12 (Bennett's Inequality)**.** *Let $X_1, X_2, \dots, X_n$ be independent real-valued random variables with zero mean, and assume that $|X_i| \leq c$ with probability one. Let*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n Var(X_i).$$

Then for any $t > 0$,

$$P\{\sum_{i=1}^n X_i > t\} \leq exp(-\frac{n\sigma^2}{c^2} h(\frac{ct}{n\sigma^2})),$$

where $h(u) = (1 + u)\log(1 + u) - u, u \geq 0$.

Theorem 13 (Bernstein's Inequality)**.** *Under the conditions of the Bennett's inequality, for any $\epsilon > 0$,*

$$P\{\frac{1}{n}\sum_{i=1}^n X_i > \epsilon\} \leq exp(-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}).$$

IV. Bounds in Vapnik-Chervonenkis Theory

A. Basic Settings

The basic components of the learning problem are (by Vapnik [9]):

1) A generator of random vectors $x \in \mathbb{R}^n$, draw independently from a fixed but unknown probability distribution function $F(x)$.
2) A supervisor (or in other words, an oracle) who returns a value y to every input x, according to a function $y = \text{oracle}(x)$.
3) A learning machine capable of implementing a set of functions $f(x, \alpha), \alpha \in \Lambda$, where Λ in fact can be a set of (even abstract) parameters.

The general setting of the learning problem can be formulated as follows.

$$\min_{\alpha \in \Lambda} R(\alpha) = \min_{\alpha \in \Lambda} \int Q(z, \alpha) dF(z), \tag{7}$$

where z describes a pair (x, y), the loss function $Q(z, \alpha)$ usually measures the discrepancy between $f(z, \alpha)$ and oracle(z), and $F(z)$ is defined on a space Z.

To minimize the risk functional (7) in practice, we have to base on finite i.i.d. samples $z_1, \dots, z_l$, which are also called the empirical data. It leads to the inductive principle of empirical risk minimization (ERM inductive principle) [9]:

1) The risk functional $R(\alpha)$ is replaced by the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^l Q(z_i, \alpha).$$

2) One approximates the function $Q(z, \alpha_0)$ that minimizes risk (7) by the function $Q(z, \alpha_l)$ that minimizes the empirical risk.

One important concept for a learning problem is called the entropy. We only consider the case where $Q(z, \alpha)$ is a set of indicator functions (for general functions, see [9]), given samples $z_1, \dots, z_l$. Consider the set of l-dimensional binary vectors

$$q(\alpha) = (Q(z_1, \alpha), \dots, Q(z_l, \alpha)).$$

When α takes various values from Λ, the vertices of the l-dimensional cube determined by $q(\alpha)$ on $z_1, \dots, z_l$ also change. The number of different vertices is defined as $N^\Lambda(z_1, \dots, z_l)$. We then define the following concepts using this number.

Definition 1. *The random entropy*

$$H^\Lambda(z_1, \dots, z_l) = \ln N^\Lambda(z_1, \dots, z_l).$$

Definition 2. *The entropy of $Q(z, \alpha)$ on samples of size l (also called the VC entropy)*

$$H^\Lambda(l) = E[\ln N^\Lambda(z_1, \dots, z_l)].$$

Definition 3. *The annealed VC entropy*

$$H^\Lambda_{ann}(l) = \ln E[N^\Lambda(z_1, \dots, z_l)].$$

Definition 4. *The growth function*

$$G^{\Lambda}(l) = \ln \sup_{z_1,\ldots,z_l} N^{\Lambda}(z_1,\ldots,z_l).$$

It is obvious that the inequalities $H^{\Lambda}(l) \leq H^{\Lambda}_{\text{ann}}(l) \leq G^{\Lambda}(l)$ hold. Note both the VC entropy and the annealed VC entropy are distribution dependent, whereas the growth function is not.

The equation

$$\lim_{l\to\infty} \frac{H^{\Lambda}(l)}{l} = 0$$

is a sufficient condition for consistency of the ERM principle.

The equation

$$\lim_{l\to\infty} \frac{H^{\Lambda}_{\text{ann}}(l)}{l} = 0$$

is a sufficient condition for a fast rate of convergence.

The equation

$$\lim_{l\to\infty} \frac{G^{\Lambda}(l)}{l} = 0$$

is a necessary and sufficient condition for consistency of ERM for any probability measure. It is also the case that if this condition holds true, the rate of convergence is fast.

Theorem 14. *Any growth function either satisfies the equality*

$$G^{\Lambda}(l) = l\ln 2,$$

or is bounded by the inequality

$$G^{\Lambda}(l) \leq h(\ln\frac{l}{h}+1),$$

where h is an integer chosen in such a way that

$$G^{\Lambda}(h) = h\ln 2,\ G^{\Lambda}(h+1) < (h+1)\ln 2.$$

From this theorem we can define the VC dimension of $Q(z,\alpha), \alpha\in\Lambda$ is h ($h=\infty$ if no such an h exists). Another definition of the VC dimension is as follows [9].

Definition 5 (VC Dimension)**.** *The VC dimension of a set of indicator functions $Q(z,\alpha), \alpha\in\Lambda$ is the maximum number h of vectors $z_1,\ldots,z_l$ that can be separated into two classes in all 2^h possible ways using functions of the set. Or in other words, the VC dimension is the maximum number of vectors that can be shattered by the set of functions.*

Definition 6 (VC Dimension)**.** *The VC dimension of a set of real functions $Q(z,\alpha), \alpha\in\Lambda$ is the VC dimension of the set of corresponding indicator functions*

$$I(z,\alpha,\beta) = \theta\{Q(z,\alpha)-\beta\},\ \alpha\in\Lambda,\ \beta\in(A,B),$$

where

$$Q(z,\alpha)\in(A,B),\ \theta(z) = \begin{cases} 0 & \text{if } z<0,\\ 1 & \text{if } z\geq 0,\end{cases}$$

and A, B can be infinite.

If for any n there exists a set of n vectors that can be shattered by the set $Q(z,\alpha), \alpha\in\Lambda$, then the VC dimension is infinity. Finiteness of the VC dimension is also a necessary and sufficient condition for distribution-independent consistency of ERM learning machines. The following example comes from [9].

Example 1 (VC Dimension)**.** *The VC dimension of the set of linear functions*

$$Q(z,\alpha) = \sum_{p=1}^{n} a_p z_p + \alpha_0,\ \alpha_0,\ldots,\alpha_p\in(-\infty,\infty),$$

is equal to $h=n+1$.

This example is a special case where the VC dimension equals the number of free parameters, which might not be true in general.

B. Bounding Theorems

First we introduce two basic inequalities.

Theorem 15. *The following holds true:*

$$P\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right| > \varepsilon\} \leq 4exp\{(\frac{H^{\Lambda}_{ann}(2l)}{l} - \varepsilon^2)l\},$$

$$P\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)dF(z)}} > \varepsilon\} \leq 4exp\{(\frac{H^{\Lambda}_{ann}(2l)}{l} - \frac{\varepsilon^2}{4})l\}.$$

Using the fact that $H^{\Lambda}_{\text{ann}}(l) \leq G^{\Lambda}(l)$, we have

Theorem 16. *The following holds true:*

$$P\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right| > \varepsilon\} \leq 4exp\{(\frac{G^{\Lambda}(2l)}{l} - \varepsilon^2)l\},$$

$$P\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)dF(z)}} > \varepsilon\} \leq 4exp\{(\frac{G^{\Lambda}(2l)}{l} - \frac{\varepsilon^2}{4})l\}.$$

Let

$$\mathcal{E} = 4\frac{G^{\Lambda}(2l) - \ln(\eta/4)}{l},^{1}$$

where l is the number of samples, η is a probability used to describe how probably the bounds are true, and $G^{\Lambda,\mathcal{B}}(l)$ is the so called growth function.

We first consider the bounded case where $A\leq Q(z,\alpha)\leq B, \alpha\in\Lambda$ (see [9]).

Theorem 17. *The following inequalities hold with probability at least $1-\eta$ simultaneously for all functions of $Q(z,\alpha), \alpha\in\Lambda$ (including the function that minimizes the empirical risk):*

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B-A}{2}\sqrt{\mathcal{E}},$$

[1]For constructing learning machines that control the generalization ability, we can use $\mathcal{E} = 4\frac{h(\ln\frac{2l}{h}+1)-\ln(\eta/4)}{l}$, or $\mathcal{E} = 2\frac{\ln N - \ln\eta}{l}$ in cases where the set of $Q(z,\alpha), \alpha\in\Lambda$ contains a finite number N of elements

$$R_{emp}(\alpha) - \frac{B-A}{2}\sqrt{\mathcal{E}} \leq R(\alpha).$$

The following inequalities hold with probability at least $1-2\eta$ *for the function* $Q(z,\alpha_l)$ *that minimizes the empirical risk:*

$$R(\alpha_l) - \inf_{\alpha\in\Lambda} R(\alpha) \leq (B-A)\sqrt{-\frac{\ln\eta}{2l}} + \frac{B-A}{2}\sqrt{\mathcal{E}}.$$

Another case is $0 \leq Q(z,\alpha) \leq B, \alpha \in \Lambda$.

Theorem 18. *The following inequality holds with probability at least* $1-\eta$ *simultaneously for all functions* $Q(z,\alpha) \leq B, \alpha \in \Lambda$ *(including the function that minimizes the empirical risk):*

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B\mathcal{E}}{2}(1+\sqrt{1+\frac{4R_{emp}(\alpha)}{B\mathcal{E}}})$$

The following inequalities hold with probability at least $1-2\eta$ *for the function* $Q(z,\alpha_l)$ *that minimizes the empirical risk:*

$$R(\alpha_l) - \inf_{\alpha\in\Lambda} R(\alpha) \leq B\sqrt{-\frac{\ln\eta}{2l}} + \frac{B\mathcal{E}}{2}(1+\sqrt{1+\frac{4}{\mathcal{E}}}).$$

The theorems above provide bounds to the generalization ability of learning machines: what actual risk $R(\alpha_l)$ is achieved by minimizing empirical risk; how close is the actual risk to the minimal possible $\inf_\alpha R(\alpha)$.

V. Bounds in Probably Approximately Correct Learning

A. Basic Settings

The concept of learnability was first proposed by Valiant in 1984 [10]. Here I have picked an extension of it by Blumer [11]. It was also Blumer who showed that the essential condition for distribution-independent learnability is finiteness of the VC dimension [11].

Definition 7. *A concept class is a nonempty set* $C \subset 2^X$ *of concepts, where* X *is a fixed set, either finite or countably infinite,* $[0,1]^n$ *of* E^n *(Euclidean* n*-dimension space) for some* $n \geq 1$.

Definition 8. *For* $\bar{x} = (x_1,\ldots,x_m) \in X^m, m \geq 1$*, the* m*-sample of* $c \in C$ *generated by* $\bar{x}$ *is given by* $sam_c(\bar{x}) = (\langle x_1, I_c(x_1)\rangle,\ldots,\langle x_m, I_c(x_m)\rangle)$.

Definition 9. *The sample space of* C*, denoted* S_C*, is the set of all* m*-samples over all* $c \in C$ *and all* $\bar{x} = (x_1,\ldots,x_m) \in X^m$*, for all* $m \geq 1$.

Definition 10. $\mathbf{A}_{C,H}$ *denotes the set of all functions* $A : S_C \to H$*, where* H *is a set of Borel sets on* X*.* H *is the hypothesis space, of which the elements are called hypotheses. For any* $A \in \mathbf{A}_{C,H}$*, probability distribution* P *on* X*,* $c \in C$*, and* $\bar{x} \in X^m$*, the error of* A *for concept* c *on* $\bar{x}$ *(w.r.t.* P*) is given by* $error_{A,c,P}(\bar{x}) = P(c\Delta h)$*, where* $h = A(sam_c(\bar{x}))$*, and "*Δ*" means symmetric difference.*

Definition 11 (PAC Learnable)**.** *If a learning function* $A \in \mathbf{A}_{C,H}$ *exists in such a way that for all* $0 < \epsilon, \delta < 1$*, for all* $c \in C$*, for all probability distribution* P*, the following holds*

$$P\{\bar{x} \in X^{m(\epsilon,\delta)} : error_{A,c,P}(\bar{x}) > \epsilon\} \leq \delta,$$

where $m(\epsilon,\delta)$ *is the sample size function. The smallest such sample size is called the sample complexity of* A*. Note by saying* C *is learnable, we mean* C *is learnable by hypothesis space* H*.*

B. Bounding Theorems

One of the bounding theorems given by Blumer is as follows [11].

Theorem 19. *Let* C *be a nontrivial, well-behaved* [2] *concept class.*

i) C *is learnable if and only if the VC dimension of* C *is finite.*

ii) *If the VC dimension of* C *is* d*, where* $d < \infty$*, then*

a) *For* $0 < \epsilon < 1$ *and sample size at least*

$$max\left(\frac{4}{\epsilon}\ln\frac{2}{\delta}, \frac{8d}{\epsilon}\ln\frac{13}{\epsilon}\right),$$

any consistent function $A : S_C \to C$ *is a learning function for* C *and*

b) *For* $0 < \epsilon < \frac{1}{2}$ *and sample size less than*

$$max\left(\frac{1-\epsilon}{\epsilon}\ln\frac{1}{\delta}, d(1-2(\epsilon(1-\delta)+\delta))\right),$$

no function $A : S_C \to H$*, for any hypothesis space* H*, is a learning function for* C*.*

These bounds have been improved in recent years. For example, it was pointed out in [12] that the sample complexity is at most

$$\frac{1}{\epsilon(1-\sqrt{\epsilon})}\left(2d\ln\frac{6}{\epsilon} + \ln\frac{2}{\delta}\right),$$

or

$$\frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right).$$

However, these improvements may still be a loose estimate.

VI. Johnson-Lindenstrauss Lemma

In [13], Johnson and Lindenstrauss first proposed the lemma (JL Lemma), which was also geometrically described by the authors as "given n points in Euclidean space, what is the smallest $k = k(n)$ so that these points can be moved into k-dimensional Euclidean space via a transformation which expands or contracts all pairwise distances by a factor of at most $1+\varepsilon$?", whereas a more common form of the lemma now is (the following theorem and lemma come from some lecture notes by Sham Kakade and Greg Shakhnarovich)

Theorem 20 (JL Lemma)**.** *Let* $\epsilon \in (0, \frac{1}{2})$*. Let* $Q \subset \mathbb{R}^d$ *be a set of* n *points and* $k = \frac{20\log n}{\epsilon^2}$*. There exists a Lipschitz mapping* $f : \mathbb{R}^d \to \mathbb{R}^k$ *such that for all* $u, v \in Q$*:*

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2. \quad (8)$$

There are various ways to prove the JL lemma (a brief overview can be found in [14]). Although not the tightest, the one I refer to uses the following lemma:

[2] See [11]for details

Lemma 4 (Norm preservation)**.** *Let $x \in \mathbb{R}^d$. Assume that the entries in $A \subset \mathbb{R}^{k\times d}$ are sampled independently from $N(0,1)$. Then*

$$P\{(1-\epsilon)\|x\|^2 \leq \|\frac{1}{\sqrt{k}}Ax\|^2 \leq (1+\epsilon)\|x\|^2\} \geq 1-2e^{-(\epsilon^2-\epsilon^3)k/4}. \tag{9}$$

The proof of Theorem 20: Choose the mapping f in such a way that $f = \frac{1}{\sqrt{k}}Ax$, where A is a $k \times d$ matrix the entry of which is sampled i.i.d. from a Gaussian $N(0,1)$. We have

$$\begin{aligned} & P\{\exists u, v, \text{s.t. the mapping fails to satisfy Eq. (8)}\} \\ \leq & \sum_{u,v\in Q} P\{\text{s.t. the mapping fails to satisfy Eq. (8)}\} \\ \leq & 2n^2 e^{-(\epsilon^2-\epsilon^3)k/4}. \end{aligned} \tag{10}$$

If we choose $k(\epsilon)$ properly (e.g., $k \geq \frac{20\ln n}{\epsilon^2}$), the probability of (10) can be strictly smaller than 1, which means such a map that satisfies (8) always exists. ■

Note the bound for k does not guarantee that k is smaller than d (dimension reduction). It can also be seen that the requirement for A to be sampled from $N(0,1)$ can be further relaxed.

REFERENCES

[1] B. C. Levy, *Principles of signal detection and parameter estimation*. Springer, 2008, ch. 4.4, pp. 131–150.

[2] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumptions," *The Annals of Mathematical Statistics*, pp. 581–586, 1951.

[3] R. McAulay and E. M. Hofstetter, "Barankin bounds on parameter estimation," *Information Theory, IEEE Transactions on*, vol. 17, no. 6, pp. 669–676, 1971.

[4] E. Barankin, "Locally best unbiased estimates," *The Annals of Mathematical Statistics*, pp. 477–501, 1949.

[5] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *Information Theory, IEEE Transactions on*, vol. 36, no. 6, pp. 1285–1301, 1990.

[6] Z. Ben-Haim and Y. C. Eldar, "The cramér-rao bound for estimating a sparse parameter vector," *Signal Processing, IEEE Transactions on*, vol. 58, no. 6, pp. 3384–3389, 2010.

[7] G. Tang and A. Nehorai, "Lower bounds on the mean-squared error of low-rank matrix reconstruction," *Signal Processing, IEEE Transactions on*, vol. 59, no. 10, pp. 4559–4571, 2011.

[8] G. Lugosi, "Concentration-of-measure inequalities," 2004.

[9] V. Vapnik, *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

[10] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[11] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the vapnik-chervonenkis dimension," *Journal of the ACM (JACM)*, vol. 36, no. 4, pp. 929–965, 1989.

[12] D. Haussler, *Probably approximately correct learning*. University of California, Santa Cruz, Computer Research Laboratory, 1990.

[13] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, p. 1, 1984.

[14] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.