Gaussian distributions with applications

Zhenan Fan, Yang Wang

Table of Contents

1 Introduction

Let (X, Y) be a random couple, where X is an instance in a space S and Y $\in \{-1, 1\}$ is a label. In machine learning, we often call X the feature, and we call functions $h : S \to \{-1, 1\}$ as (binary) classfiers.

Let $\mathcal{G}$ be set of functions from S into $\mathbb{R}$. And sign$(g(X))$ will be used as a predictor(or classifier) of the unknown label Y. Often we would not have the distribution of (X, Y), in this case, our choice of the classifier is based on a set which we called the training data $(X_1, Y_1), ...(X_n, Y_n)$ that consists of n independent and identically distributed (ie: i.i.d) copies of (X, Y). Our goal is through learning, to find a classfier $\hat{g} \in \mathcal{G}$, such that its generalization error is small(on test sets). The standard approach to this problem is based on bounding the difference between the generalization error $\mathbb{P}(Yg(X) \leq 0)$ and the training error $\frac{1}{n}\sum_{j=1}^{n} I_{\{Y_j g(X_j) \leq 0\}}$ uniformly over the whole class $\mathcal{G}$. We can define $f(X) = Yg(X)$, so mislabel happens when $f(X) \leq 0$. In theorem 2, we develop an upper bound for the probability of there existing a function f in F, such that $P(f \leq 0)$ greater than some quantity associated with the Gaussian complexity function of class $\mathcal{F}$ and positive t is bounded by some quantity in terms of t which we can easily compute.

In section 3, we will introduce some basic information about neural network learning, including forward inference and backward update. And in section 4, we would show how to bound generalization errors in neural network learning. In theorem 3, we develop a probabilistic bound for such error, which is a specific extension of theorem 2 with some new setting associated with neural network.

Finally, in section 5 and 6, in order to make the bound more clear, we choose some specific $\mathcal{H}$, which is half space, as the space of base functions in neural network learning, and construct a bound for Gaussian Complexity term $G_n(\mathcal{H})$, by using Vapnik-Chervonenkis dimension.The result is very convincing, we showed that the Gaussian Complexity term will go to zero as the size of training set increases.

2 Probabilistic bounds for general function classes in terms of Gaussian and Rademacher complexities

Let $(S, \mathcal{A}, P)$ be a probability space and let $\mathcal{F}$ be a class of measurable function from $(S, \mathcal{A})$ into $\mathbb{R}$. We could also replace S by $S \times \{-1, 1\}$ later when we want to talk about classification problems. Let $\{X_k\}$ be a sequence of i.i.d random variables taking values in $(S, \mathcal{A})$ with common distribution P. We assume that this sequence is defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. Let P_n be the empirical measure based on the sample $(X_1, ...X_n)$.

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(X_i).$$

Furthermore, P_n can map measurable function f to its empirical mean:

$$f \mapsto P_n f = \int_S f dP_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Let $L^\infty(\mathcal{F})$ denotes the Banach space of uniformly bounded real valued function on $\mathcal{F}$ with the norm

$$\|Y\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |Y(f)|.$$

Our goal is to find bounds on $P(f \leq 0)$ and on the difference $|P_n(f \leq 0) - P(f \leq 0)|$ that holds for all $f \in \mathcal{F}$ with high probability.

We define Gaussian complexity function of the class $\mathcal{F}$

$$G_n(\mathcal{F}) := \mathbb{E} \sup_{f \in \mathcal{F}} |n^{-1} \sum_{i=1}^{n} g_i f(X_i)|.$$

And the Rademacher complexity function

$$R_n(\mathcal{F}) := \mathbb{E} \sup_{f \in \mathcal{F}} |n^{-1} \sum_{i=1}^{n} \epsilon_i f(X_i)|,$$

where $\{\epsilon_i\}$ is a sequence of i.i.d random variable(taking values $+1$ or -1 with probability both $1/2$). Here we assume that for any $x \in S$ the set of real numbers $\{f(x) : f \in \mathcal{F}\}$ is bounded.

2.1 Martingale Difference Inequality

We want to estimate $P(f \leq 0)$. Define such a Lipschitz function such that $\varphi(x) \geq I_{(-\infty,0]}$. We have $P(f \leq 0) \leq P_n\varphi(f) + \|P_n - P\|_{\mathcal{G}_\varphi}$, where $G_\varphi := \{\varphi \circ f : f \in \mathcal{F}\}$. Since $P_n\varphi(f)$ is what we have known based on our data. We want to bound $\|P_n - P\|_{\mathcal{G}_\varphi}$. One way to do this is to find how far it is from its expectation. Let $\|P_n - P\|_{\mathcal{G}_\varphi} = Z$, we actually want to find a bound for $\mathbb{P}(Z - \mathbb{E}Z > t)$, for any $t \geq 0$.

Let $Z(x_1, ...x_n) : \mathcal{X} \mapsto \mathbb{R}$. We would like to bound $Z - \mathbb{E}Z$. We make the following assumptions. For any $x_1, ...x_n, x'_1, x'_2.....x'_n$,

$$|Z(x_1, ..x_i, ...x_n) - Z(x_1, ...x_{i-1}, x'_i, x_{i+1}, ...x_n)| \leq c_i$$

for some constant c_i. We can decompose $Z - \mathbb{E}Z$ as follows

$$\begin{aligned} Z(x_1, ...x_n) - E_{x'}Z(x'_1, ...x'_n) &= (Z(x_1, ...x_n) - E_{x'}Z(x'_1, x_2, ..., x_n)) + \\ &\quad (E_{x'}Z(x'_1, x_2, ...x_n) - E_{x'}Z(x'_1, x'_2, x_3....x_n)) \\ &\quad + (E_{x'}Z(x'_1, , x'_{n-1}...x_n) - E_{x'}Z(x'_1, x'_2, ...x'_n)) \\ &= Z_1 + Z_2 + Z_3.....Z_n, \end{aligned}$$

where $Z_i = \mathbb{E}_{x'}Z(x'_1, .., x'_{i-1}, x_i..x_n) - \mathbb{E}_{x'}(x'_1, ..x'_i, x_{i+1}....x_n)$.

Assume $(1)|Z_i| \leq c_i$ $(2)E_{X_i}Z_i = 0$ $(3)Z_i = Z_i(x_1, ...x_n)$

Lemma 1 *For any $\lambda \in \mathbb{R}$,*

$$E_{x_i}e^{\lambda Z_i} \leq e^{\lambda^2 c_i^2}/2$$

Proof. Take any $-1 \leq s \leq 1$. Since function $e^{\lambda s}$ is convex and

$$e^{\lambda s} = e^{\lambda(\frac{1+s}{2}) + (-\lambda)(\frac{1-s}{2})}.$$

Check that $0 \leq \frac{1+s}{2}, \frac{1-s}{2} \leq 1$, and $\frac{1+s}{2} + \frac{1-s}{2} = 1$, hence we have

$$e^{\lambda s} \leq \frac{1+s}{2}e^{\lambda} + \frac{1-s}{2}e^{-\lambda} = \frac{e^{\lambda} + e^{-\lambda}}{2} + s\frac{e^{\lambda} - e^{-\lambda}}{2} \leq e^{\lambda^2/2} + s \cdot \sinh(\lambda)$$

by Taylor expansion. Now use $\frac{Z_i}{c_i} = s$, and since by assumptions, $-1 \leq \frac{Z_i}{c_i} \leq 1$. We have

$$e^{\lambda Z_i} = e^{\lambda c_i \cdot \frac{Z_i}{c_i}} \leq e^{\lambda^2 c_i^2/2} + \frac{Z_i}{c_i}\sinh(\lambda c_i)$$

since $\mathbb{E}_{x_i}Z_i = 0$ We now have $E_{x_i}e^{\lambda Z_i} \leq e^{\lambda^2 c_i^2/2}$.

□

Lemma 2 *If condition of lemma 1 is satisfied for each i, we have*

$$\mathbb{P}(Z - \mathbb{E}Z > t) \leq e^{\frac{t^2}{2\sum_{i=1}^n c_i^2}}.$$

Proof. For any $\lambda > 0$

$$\mathbb{P}(Z - \mathbb{E}Z > t) = \mathbb{P}(e^{\lambda(Z-\mathbb{E}Z)} > e^{\lambda t}) \leq \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}}$$

Since we have

$$\begin{aligned}
\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} &= \mathbb{E}e^{\lambda(Z_1+Z_2+....Z_n)} \\
&= \mathbb{E}\mathbb{E}_{x_1}e^{\lambda(Z_1+Z_2+....Z_n)} \\
&= \mathbb{E}[e^{\lambda(Z_2+....Z_n)}\mathbb{E}_{x_1}e^{\lambda Z_1}] \\
&\leq \mathbb{E}[e^{\lambda(Z_2+....Z_n)}\mathbb{E}_{x_1}e^{\lambda^2 c_i^2/2}] \\
&= e^{\lambda^2 c_1^2/2}\mathbb{E}\mathbb{E}_{x_2}[e^{\lambda(Z_2+....Z_n)}] \\
&\leq e^{\lambda^2(c_1^2+c_2^2)/2}\mathbb{E}e^{\lambda(Z_3+....Z_n)} \\
&\leq e^{\lambda^2\sum_{i=1}^n c_i^2/2}.
\end{aligned}$$

Hence

$$\mathbb{P}(Z - \mathbb{E}Z > t) \leq e^{-\lambda t + \lambda^2\sum_{i=1}^n c_i^2/2}.$$

We minimize the exponent of the right hand side with respect to λ. We have $\lambda = \dfrac{t}{\sum_{i=1}^n c_i^2}$. Substitute it back into the inequality we can get our result. □

Lemma 3 *Let $\mathbb{F}$ be a class of functions: $\mathcal{X} \mapsto [a,b]$. Define the empirical process*

$$Z(x_1,..x_n) = \sup_{f\in\mathcal{F}} |\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)|.$$

We have

$$\mathbb{P}(Z - \mathbb{E}Z > t) \leq e^{-\frac{nt^2}{2(b-a)^2}}.$$

Proof. First we show that, for any i,

$$\begin{aligned}
|Z(x_1,..x_i',...x_n) - Z(x_1,..x_i,...x_n)| &= |\sup_f |\mathbb{E}f - \frac{1}{n}(f(x_1) + ..f(x_i') + ..f(x_n))| \\
&\quad - \sup_f |\mathbb{E}f - \frac{1}{n}(f(x_1) + ..f(x_i) + ..f(x_n))|| \\
&\leq \sup_{f\in\mathcal{F}} \frac{1}{n}|f(x_i) - f(x_i')| \\
&\leq \frac{b-a}{n}.
\end{aligned}$$

We set $c_i = \frac{b-a}{n}$ for all i, then by lemma 2,

$$\mathbb{P}(Z - \mathbb{E}Z > t) \leq \exp(-\frac{t^2}{2\sum_{i=1}^n \frac{(b-a)^2}{n^2}}) = e^{-\frac{nt^2}{2(b-a)^2}}.$$

□

2.2 Symmetrization

As we have found a bound for $\mathbb{P}\{\|P_n - P\|_{\mathcal{G}_\varphi} - \mathbb{E}\|P_n - P\|_{\mathcal{G}_\varphi} \geq t\}$, we still want to bound $\mathbb{E}\|P_n - P\|_{\mathcal{G}_\varphi}$ in terms of Rademacher complexity function and Gaussian complexity function.

Lemma 4 *Denote* $\mathbb{P}_n^o f = \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)$.

$$\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}} \leq 2\mathbb{E}\|\mathbb{P}_n^o\|_{\mathcal{F}}.$$

Proof. Here we introduce copies of X_i, X_i', each of them is independent with each other.

$$\begin{aligned}
E\sup_f |\frac{1}{n}\sum_{i=1}^n f(X_i) - Ef| &= E\sup_f \frac{1}{n}|\sum(f(X_i) - Ef(X_i'))| \\
&= E_{X_i}\sup_f |\frac{1}{n}E_{X_i'}\sum(f(X_i') - f(X_i))| \\
&\leq E_{X_i}\sup_f \frac{1}{n}E_{X_i'}|\sum(f(X_i') - f(X_i))| \\
&\leq E_{X_i}E_{X_i'}\sup_f |\frac{1}{n}\sum_{i=1}^n(f(X_i) - f(X_i'))|.
\end{aligned}$$

Since $f(X_i)$ and $f(X_i')$ are identical copies, so the we can introduce Rademacher random variable ϵ_i that is independent to X_i and X_i'. So by the symmetry of Rademacher variable and triangle inequality, we have

$$E\sup_f |\frac{1}{n}\sum_{i=1}^n f(X_i) - Ef| \leq 2E\frac{1}{n}\sup|\sum_{i=1}^n \epsilon_i f(X_i)|.$$

As desired.

□

Lemma 5

$$\mathbb{E}\|\mathbb{P}_n^o\|_{\mathcal{F}} \leq \sqrt{2\pi}\mathbb{E}\sup_{h\in\mathcal{F}}|n^{-1}\sum_{i=1}^n g_i h(X_i)|.$$

Proof. We can start from last lemma's last step. Consider Radmacher variable ϵ_i, it is the same as $\frac{g_i}{\|g_i\|}$, where g_i is a standard gaussian. So if we replace ϵ_i in our previous equation. We have

$$E\sup_f |\frac{1}{n}\sum_{i=1}^n f(X_i) - Ef| \leq \frac{2}{E\|g_1\|}E\frac{1}{n}\sup|\sum_{i=1}^n g_i f(X_i)|.$$

And since $E\|g_1\|$ is $\sqrt{\frac{\pi}{2}}$, we have our result.

□

Lemma 6 *(Rademacher comparison inequality) Let F:* $\mathbb{R}_+ \to \mathbb{R}_+$ *be convex and increasing. Let further* $\varphi_i : \mathbb{R} \to \mathbb{R}$, *be contractions such that* $\varphi_i(0) = 0$. *Then for any bounded subset T in* $\mathbb{R}^N$, *we have*

$$\mathbb{E}F(\frac{1}{2}\|\sum_{i=1}^N \epsilon_i\varphi_i(t_i)\|_T) \leq \mathbb{E}F(\|\sum_{i=1}^N \epsilon_i t_i\|_T).$$

We would not prove this theoreom in our project. Readers can refer to this book: *Probability in Banach Spaces ,Ledoux, Michel, Talagrand, Michel*, and it is the theorem 4.1.2

Now, we prove our first theorem by putting above pieces together.

2.3 Probabilistic bounds for general function classifiers

Theorem 1 *For all $t > 0$,*

$$\mathbb{P}\left\{\exists f \in \mathcal{F} : P\{f \leq 0\} > \inf_{k\geq 1}[P_n\varphi_k(f) + 4L(\varphi_k)R_n(\mathcal{F}) + (\frac{2\log k}{n})^{1/2}] + \frac{t}{\sqrt{n}}\right\} \leq 2\exp(-t^2/2)$$

and

$$\mathbb{P}\left\{\exists f \in \mathcal{F} : P\{f \leq 0\} > \inf_{k\geq 1}[P_n\varphi_k(f) + \sqrt{2\pi}L(\varphi_k)G_n(\mathcal{F}) + (\frac{2\log k}{n})^{1/2}] + \frac{t+2}{\sqrt{n}}\right\} \leq 2\exp(-t^2/2)$$

Proof. Without loss of generality we can assume that each φ takes values in $[0, 1]$. In this case we have $\varphi(x) = 1$ for $x \leq 0$. Fix $\varphi \in \Phi$, for all $f \in \mathcal{F}$ we have

$$P\{f \leq 0\} \leq P\varphi(f) = P\varphi(f) - P_n\varphi(f) + P_n\varphi(f) \leq P_n\varphi(f) + \|P_n - P\|_{\mathcal{G}_\varphi}. \tag{1}$$

By the definition of our norm. Here

$$\mathcal{G}_\varphi := \{\varphi \circ f - 1 : f \in \mathcal{F}\}$$

By lemma 3, and substitute t with $\frac{t}{\sqrt{n}}$. Since function in $\mathcal{G}_\varphi$ have range $[-1, 0]$, we have

$$\mathbb{P}\{\|P_n - P\|_{\mathcal{G}_\varphi} \geq \mathbb{E}\|P_n - P\|_{\mathcal{G}_\varphi} + \frac{t}{\sqrt{n}}\} \leq \exp(-t^2/2).$$

Thus with probability at least $1 - \exp(t^2/2)$ for all $f \in \mathcal{F}$

$$P(f \leq 0) \leq P_n\varphi(f) + \mathbb{E}\|P_n - P\|_{\mathcal{G}_\varphi} + \frac{t}{\sqrt{n}}. \tag{2}$$

And by Symmetrization lemma (lemma 4)

$$\mathbb{E}\{\|P_n - P\|_{\mathcal{G}_\varphi}\} \leq 2\mathbb{E}\|\mathbb{P}_n^o\|_{\mathcal{G}_\varphi} \tag{3}$$

Since a function $\frac{\varphi - 1}{L(\varphi)}$ is a contraction and $\varphi(0) - 1 = 0$, the Rademacher comparison inequality implies

$$\mathbb{E}_\epsilon \sup_{h\in\mathcal{G}_\varphi} |\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h(X_i)| \leq 2L(\varphi)\mathbb{E}_\epsilon \sup_{h\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)|$$

Where $h = \varphi \circ f - 1$. Then it follows from (2),(3) that with probability at least $1 - e^{-t^2/2}$ we have for all $f \in \mathcal{F}$

$$P\{f \leq 0\} \leq P_n\varphi(f) + 4L(\varphi)R_n(\mathcal{F}) + \frac{t}{\sqrt{n}} \tag{4}$$

We can now use (4) with $\varphi = \varphi_k$ and t replaced by $t + 2\sqrt{\log k}$ and obtain

$$\begin{aligned}&\mathbb{P}\Big\{\exists f\in\mathcal{F}:P\{f\le 0\}>\inf_{k\ge 1}[P_n\varphi_k(f)+4L(\varphi_k)R_n(\mathcal{F})+\sqrt{2}(\frac{\log k}{n})^{1/2}]+\frac{t}{\sqrt{n}}\Big\}\\&\le\sum_{k\ge 1}\exp\{-(t+2\sqrt{\log k})^2/2\}\le\sum_{k\ge 1}k^{-2}e^{-t^2/2}=\frac{\pi^2}{2}e^{-t^2/2}\le 2e^{-t^2/2}\end{aligned}\tag{5}$$

The proof for the second bound is similar with some changes. We define class $\mathcal{G}_\varphi$ as $\{\varphi\circ f:f\in\mathcal{F}\}$. For further develop the inequality in (3). And we have

$$\mathbb{E}\|P_n-P\|_{\mathcal{G}_\varphi}\le 2\mathbb{E}\|\mathbb{P}_n^o\|_{\mathcal{G}_\varphi}\le\sqrt{2\pi}\mathbb{E}\sup_{h\in\mathcal{G}_\varphi}|n^{-1}\sum_{i=1}^n g_ih(X_i)|\tag{6}$$

The second inequality is by the lemma 5.

Define these two Gaussian processes:

$$Z_1(f,\sigma):=\sigma n^{-1/2}\sum_{i=1}^n g_i(\varphi\circ f)(X_i)$$

$$Z_2(f,\sigma):=L(\varphi)n^{-1/2}\sum_{i=1}^n g_if(X_i)+\sigma g$$

where $\sigma=\pm 1$, and g is a standard Gaussian independent of g_i. We can denote by $\mathbb{E}_g$ the expectation on the probability space $(\Omega_g,\Sigma_g,\mathbb{P}_g)$ on which the g_i and g are defined. Then we have

$$\mathbb{E}_g|Z_1(f,\sigma)-Z_1(h,\sigma')|^2\le\mathbb{E}_g|Z_2(f,\sigma)-Z_2(h,\sigma')|^2$$

We will prove the above inequality in two cases. First, if we have $\sigma\sigma'=1$. Then

$$|Z_1(f,\sigma)-Z_1(h,\sigma')|^2=1/n|\sum_{i=1}^n g_i(\varphi(f(X_i))-\varphi(h(X_i))|^2$$

Since each g_i is independent. If we take expectation, we would have

$$\begin{aligned}\mathbb{E}_g|Z_1(f,\sigma)-Z_1(h,\sigma')|^2&=1/n\sum_{i=1}^n|\varphi(f(X_i))-\varphi(h(X_i))|^2\\&\le L(\varphi)^2/n\sum_{i=1}^n(f(X_i)-h(X_i))^2\\&=\mathbb{E}_g|Z_2(f,\sigma)-Z_2(h,\sigma')|^2\end{aligned}$$

If $\sigma\sigma'=-1$.

$$\begin{aligned}\mathbb{E}_g|Z_1(f,\sigma)-Z_1(h,\sigma')|^2&=1/n\sum_{i=1}^n|\varphi(f(X_i))-\varphi(h(X_i))|^2\\&\le 2n^{-1}\sum_{i=1}^n\varphi^2(f(X_i))+2n^{-1}\sum_{i=1}^n\varphi^2(h(X_i))\\&\le\mathbb{E}(2g)^2\le\mathbb{E}_g|Z_2(f,\sigma)-Z_2(h,\sigma')|^2\end{aligned}$$

By slepian's lemma in lecture notes, we have

$$\mathbb{E}_g\sup\{Z_1(f,\sigma):f\in\mathcal{F},\sigma=\pm 1\}\le\mathbb{E}_g\sup\{Z_2(f,\sigma):f\in\mathcal{F},\sigma=\pm 1\}$$

We have

$$\mathbb{E}_g \sup\{Z_1(f,\sigma) : f \in \mathcal{F}, \sigma = \pm 1\} = \mathbb{E}_g \sup_{h \in \overline{\mathcal{G}}_\varphi} [n^{-1/2} \sum_{i=1}^n g_i h(X_i)]$$

Where we define $\overline{\mathcal{G}}_\varphi := \{\varphi(f), -\varphi(f) : f \in \mathcal{F}\}$ And it is easy to see that

$$L(\varphi)\mathbb{E}_g \sup |n^{-1/2} \sum_{i=1}^n g_i f(X_i)| + \mathbb{E}|g| \geq \mathbb{E}_g \sup\{Z_2(f,\sigma) : f \in \mathcal{F}, \sigma = \pm 1\}$$

So we have

$$\mathbb{E}_g \sup_{h \in \overline{\mathcal{G}}_\varphi} [n^{-1/2} \sum_{i=1}^n g_i h(X_i)] \leq L(\varphi)\mathbb{E}_g \sup |\sum_{i=1}^n g_i f(X_i)| + n^{-1/2}\mathbb{E}|g| \tag{7}$$

So by (2), (6), and (7), we have that with probability at least $1 - \exp(t^2/2)$ for all $f \in \mathcal{F}$,

$$P(f \leq 0) \leq P_n\varphi(f) + \mathbb{E}\|P_n - P\|_{\mathcal{G}_\varphi} + \frac{t}{\sqrt{n}} \leq P_n\varphi(f) + \sqrt{2\pi}L(\varphi)G_n(\mathcal{F}) + \frac{t+2}{\sqrt{n}}. \tag{8}$$

Then by the same method we used for the first bound we can derive the one for the second one. □

Now let us consider a special family of functions for φ. Let

$$\Phi_0 := \{\varphi(./\delta) : \delta \in (0,1]\},$$

we can check that $L(\varphi(./\delta)) \leq L(\varphi)/\delta$.

Theorem 2 *For all $t > 0$,*

$$\mathbb{P}\left\{\exists f \in \mathcal{F} : P\{f \leq 0\} > \inf_{\delta \in (0,1]} [P_n\varphi(\frac{f}{\delta}) + \frac{8L(\varphi)}{\delta} R_n(\mathcal{F}) + (\frac{2\log\log_2(2\delta^{-1})}{n})^{1/2}] + \frac{t}{\sqrt{n}}\right\} \leq 2\exp(-t^2/2)$$

and

$$\mathbb{P}\{\exists f \in \mathcal{F} : P\left\{f \leq 0\} > \inf_{\delta \in (0,1]} [P_n\varphi(\frac{f}{\delta}) + \frac{2\sqrt{2\pi}L(\varphi)}{\delta} G_n(\mathcal{F}) + (\frac{2\log\log_2(2\delta^{-1})}{n})^{1/2}] + \frac{t+2}{\sqrt{n}}\right\} \leq 2\exp(-t^2/2)$$

Proof.

We can apply the bounds of Theorem 1 for choosing a sequence of $\varphi_k(.) := \varphi(./\delta_k)$, where $\delta_k = 2^{-k}$. And notice that for $\delta \in (\delta_k, \delta_{k-1}]$, we have

$$\frac{1}{\delta_k} \leq \frac{2}{\delta}, P_n\varphi(\frac{f}{\delta_k}) \leq P_n\varphi(\frac{f}{\delta})$$

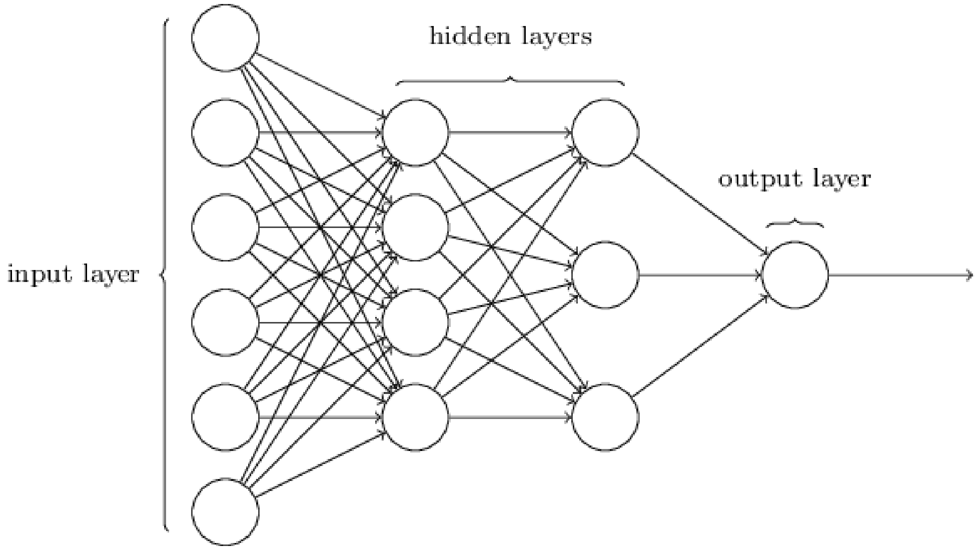
and

$$\sqrt{\log k} = \sqrt{\log\log_2 \frac{1}{\delta_k}} \leq \sqrt{\log\log_2 \frac{1}{\delta}}$$

□

3 Neural Network Learning

Many machine learning methods are inspired by biology, brains and Neural Network is a popular supervised approach. Artificial neurons are called units and each unit computes its value based on linear combination of values of units that point into it, which is called forward pass.



Network.png

We can take a neural network with one hidden layer as an example. Suppose we have D inputs, J hidden units and K outputs. And we use h_j's to denote the hidden units and O_k's to denote the outputs.Then the output of network can be written as

$$h_j(x) = f(v_{j0} + \sum_{i=1}^{D} v_{ji} x_i)$$

and

$$O_k(x) = g(w_{k0} + \sum_{j=1}^{J} w_{kj} h_j(x)).$$

And then we use backward propagation to update the weights (via gradient descent). To be more specific, we choose a suitable error function E, and then we calculate the derivative with respect to each weights, namely $\frac{\partial E}{\partial w_{ki}}$. And then we update the weights by gradient descent, namely

$$w_{ki} \leftarrow w_{ki} - \beta \frac{\partial E}{\partial w_{ki}}.$$

In our project, we try to bound the generalization error $P(f < 0)$ by

$$P(f < 0) \leq P_n(f < 0) + \sup_{f \in \mathcal{F}} |P(f < 0) - P_n(f < 0)|,$$

and by the characteristics of gradient descent, we know that the emprical error $P_n(f < 0)$ can be regarded as a small term, and thus our main theorem is more about the bound of $\sup_{f \in \mathcal{F}} |P(f < 0) - P_n(f < 0)|$.

4 Bounding the generalization error in neural network learning

Let $\mathcal{H}$ be a class of measurable functions from $(\S, \mathcal{A})$ into $\mathbb{R}$ and let G be an arbitrary acyclic graph with a unique input vertex v_i and a unique output vertex v_o.

Suppose the set V of all neurons is divided in to layers (based on different depth)

$$V = \{v_i\} \cup \bigcup_{j=0}^{l} V_j,$$

where $V_l = \{v_o\}$. And the neurons in V_0 will be called base neurons.

Suppose that the inputs of the neurons of the layer $V_j, j \geq 1$ are the ouputs of the neurons from V_{j-1}. And the inputs of V_0 are the outputs of v_i. Then we assign the labels to the neurons of the network:

- Each neuron of V_0 is labeled by a function from the base class $\mathcal{H}$.
- Each neuron of the jth layer V_j , where $j \geq 1$, is labeled by a vector

$$w = (w_1, \ldots, w_n) \in \mathbb{R}^n,$$

where n is the number of inputs of the neuron and w_i is the weight for ith input.

Given a Borel function σ from $\mathbb{R}$ to $[-1, 1]$ and a vector $w = (w_1, \ldots, w_n)$, we can define a function $N_{\sigma,w} : \mathbb{R}^n \to \mathbb{R}$ by

$$N_{\sigma,w}(u_1, \ldots, u_n) = \sigma(\sum_{i=1}^{n} w_i u_i),$$

and let $\{\sigma_j : j \geq 1\}$ be the set of functions from $\mathbb{R}$ to $[-1, 1]$, satisfying the Lipschitz conditions:

$$|\sigma_j(u) - \sigma_j(v)| \leq L_j |u - v|.$$

Then given an instance $x \in \mathcal{S}$, the network works the following way:

- A base neuron computes the value of the labeled base function on x, and outputs the value through its output edges. Namely assume the neuron is labeled by $h \in \mathcal{H}$ and take input x, then the output is $h(x)$.
- For a neuron in the jth layer, $j \geq 1$, assume it is labeled by $w = (w_1, \ldots, w_n)$ and has input $u = (u_1, \ldots, u_n)$. Then the corresponding output is $N_{\sigma_j,w}(u)$.

Now we define class of functions computable by neural network with weights bounded. Given $\{A_j : j \geq 0\}$ as sequence of positive numbers, then we recursively define the classes by:

$$\mathcal{H}_0 = \mathcal{H},$$

$$\mathcal{H}_j = \{N_{\sigma_j,w}(h_1, \ldots, h_n) : n \geq 0, h_i \in \mathcal{H}_{j-1}, w \in \mathbb{R}^n, \sum_{i=1}^{n} |w_i| \leq A_j\}, \text{ for } j \geq 1.$$

Also let $\mathcal{H}_\infty = \bigcup_{j=0}^{\infty} \mathcal{H}_j$: all functions computable by neural network with base $\mathcal{H}$ and bounded weights.

Theorem 3 *For all $t > 0$ and all $l \geq 1$,*

$$\mathbb{P}\big\{\exists f \in \mathcal{H}_l : P(\tilde{f} \leq 0) > \inf_{\delta \in [0,1]} \big[P_n \varphi(\frac{\tilde{f}}{\delta}) + \frac{2\sqrt{2\pi} L(\varphi)}{\delta} \prod_{j=1}^{l} L_j A_j G_n(\mathcal{H}) + (\frac{\log \log_2(2\delta^{-1})}{n})^{\frac{1}{2}}\big] + \frac{t+2}{\sqrt{n}}\big\} \leq 2\exp(-\frac{1}{2}t^2),$$

where $\tilde{f} = f \times label$.

Proof. Applying Theorem 2 to class $\mathcal{F} = \mathcal{H}_l$, we will have that for all $t > 0$

$$\mathbb{P}\{\exists f \in \mathcal{H}_l : P(\tilde{f} \leq 0) > \inf_{\delta\in[0,1]} \left[P_n\varphi(\frac{\tilde{f}}{\delta}) + \frac{2\sqrt{2\pi}L(\varphi)}{\delta}G_n(\mathcal{H}_l) + (\frac{\log\log_2(2\delta^{-1})}{n})^{\frac{1}{2}}\right] + \frac{t+2}{\sqrt{n}}\} \leq 2\exp(-\frac{1}{2}t^2).$$

So now in order to prove the theorem, we only need to show that

$$G_n(\mathcal{H}_l) = \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{H}_l} \leq \prod_{j=1}^{l} L_jA_j\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{H}} = \prod_{j=1}^{l} L_jA_jG_n(\mathcal{H}).$$

Define

$$\mathcal{H}'_l = \{\sum_{i=1}^{n} w_ih_i : n \geq 0, hi \in \mathcal{H}_{l-1}, w \in \mathbb{R}^n, \|w\|_{l_1} \leq A_l\}$$

which is actually $\mathcal{H}_l$ without applying the function σ_l.

Then for $f \in \mathcal{H}'_l$, we consider two Gaussian processes:

$$Z_1(f) = n^{-\frac{1}{2}}\sum_{i=1}^{n} g_i(\sigma_l \circ f)(X_i)$$

and

$$Z_2(f) = L_ln^{-\frac{1}{2}}\sum_{i=1}^{n} g_if(X_i)).$$

Then we have

$$\begin{aligned}
& \mathbb{E}_g|Z_1(f) - Z_1(h)|^2 \\
= \;& \mathbb{E}_g|n^{-\frac{1}{2}}\sum_{i=1}^{n} g_i(\sigma_l \circ f)(X_i) - n^{-\frac{1}{2}}\sum_{i=1}^{n} g_i(\sigma_l \circ h)(X_i)|^2 \\
= \;& n^{-1}\mathbb{E}_g|\sum_{i=1}^{n} g_i(\sigma_l(f(X_i)) - \sigma_l(h(X_i)))|^2 \\
= \;& n^{-1}\sum_{i=1}^{n} |\sigma_l(f(X_i)) - \sigma_l(h(X_i))|^2 \\
\leq \;& L_l^2n^{-1}\sum_{i=1}^{n} |f(X_i) - h(X_i)|^2 \\
= \;& \mathbb{E}_g|Z_2(f) - Z_2(h)|^2 \qquad (9)
\end{aligned}$$

Lemma 7 *Let $X = (X_i)_{i\leq n}$ and $Y = (Y_i)_{i\leq n}$ be two Gaussian vectors in $\mathbb{R}^n$ such that*

$$\mathbb{E}(X_i - X_j)^2 \geq \mathbb{E}(Y_i - Y_j)^2 \text{ for all } i, j \leq n.$$

Then, the Slepian's inequality holds, i.e.

$$\mathbb{E}\max_i X_i \geq \mathbb{E}\max_i Y_i.$$

Proof. The proof for this lemma is in the lecture notes. □

Then by lemma 7 and equation (9), we can conclude that

$$\mathbb{E}_g \|Z_1\|_{\mathcal{H}_l'} \quad \leq \quad \mathbb{E}_g \|Z_2\|_{\mathcal{H}_l'} \tag{10}$$

And it is obvious that

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_l} = n^{-\frac{1}{2}}\mathbb{E}_g \|Z_1\|_{\mathcal{H}_l'}$$

and

$$L_l \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_l'} = n^{-\frac{1}{2}}\mathbb{E}_g \|Z_2\|_{\mathcal{H}_l'}.$$

So it follows that

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_l} \leq L_l \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_l'}.$$

Definition. A symmetric convex hull of A is the smallest convex set including A and $-A$ and is the set of all finite convex combination

$$\sum \lambda_i a_i$$

with $a_i \in A \cup -A$, $\sum \lambda_i \leq 1$.And the closed symmetric convex hull is just the closure of the symmetric convex hull.

Then obviously, we can see that $\mathcal{H}_l' = A_l conv_s(\mathcal{H}_{l-1})$, where $conv_s(\mathcal{H}_{l-1})$ denotes the closed symmetric convex hull of the class $\mathcal{H}_{l-1}$. Then it follows that

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_l'} \leq A_l \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_{l-1}}. \tag{11}$$

Now combining with equation (11), we can get

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_l'} \leq L_l A_l \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_{l-1}'}$$

Then by simple induction, we can conclude that

$$G_n(\mathcal{H}_l') = \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}_l'} \leq \prod_{j=1}^{l} L_j A_j \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{H}} = \prod_{j=1}^{l} L_j A_j G_n(\mathcal{H}).$$

□

5 Vapnik-Chervonenkis classes of sets

Definition. Assume $\mathcal{C}$ is a class of sets and for any sample $\{x_1, \ldots, x_n\}$, let

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = card\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}$$

and

$$\Delta_n(\mathcal{C}) = \sup_{\{x_1,\ldots,x_n\}} \leq \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq 2^n.$$

Note that if $\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = 2^n$, then we say $\mathcal{C}$ shatters $\{x_1, \ldots, x_n\}$. If there exist some constant $V < \infty$ such that

$$\begin{cases} \Delta_n(\mathcal{C}) = 2^n & n \leq V, \\ \Delta_n(\mathcal{C}) < 2^n & n > V. \end{cases}$$

then $\mathcal{C}$ is called a VC class and V is called VC dimension of $\mathcal{C}$.

Lemma 8 ***Sauer's lemma***
$\forall\{x_1, \ldots, x_n\}$,

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq \left(\frac{en}{V}\right)^V$$

where V is the VC dimension of $\mathcal{S}$.

Proof. First, we show that $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ is bounded by the number of subsets shattered by $\mathcal{C}$. Without loss of generality, we can replace $\mathcal{C}$ by $\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}$.

We will say that C is hereditary if $B \subset C \in \mathcal{C} \Rightarrow B \in \mathcal{C}$. Apparently, if $\mathcal{C}$ is hereditary, then every $C \in \mathcal{C}$ is shattered by $\mathcal{C}$, and the lemma follows.

Now if $\mathcal{C}$ is not hereditary, then we will build a $\mathcal{C}'$ such that $\mathcal{C}'$ has the same cardinality as $\mathcal{C}$ and the number of shattered subsets is not increasing.

Define operator T_i for $i = 1, \ldots, n$ by

$$T_i(C) = \begin{cases} C - \{x_i\} & C - \{x_i\} \notin \mathcal{C}, \\ C & otherwise. \end{cases}$$

and

$$T_i(\mathcal{C}) = \{T_i(C) : C \in \mathcal{C}\}.$$

It follows that $|T_i(\mathcal{C})| = |\mathcal{C}|$ and $\forall A \subset \{x_1, \ldots, x_n\}$, if A is shattered by $T_i(\mathcal{C})$, then A is also shattered by $\mathcal{C}$.

Then apply $T = T_1 \circ \cdots \circ T_n$ until $T^{k+1}(\mathcal{C}) = T^k(\mathcal{C})$. Since if $T_i(\mathcal{C}) \neq \mathcal{C}$, then $\sum_{C \in \mathcal{C}} |T_i(C)| < \sum_{C \in \mathcal{C}} |C|$, it follows the end condition will be met after at most $\sum_{C \in \mathcal{C}} |C|$ times. And apparently the resulting $\mathcal{C}'$ is hereditary, and then $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ is bounded by the number of subsets shattered by $\mathcal{C}$.

Next, we show that

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq \left(\frac{en}{V}\right)^V.$$

$$\begin{aligned}\Delta_n(\mathcal{C}, x_1, \ldots, x_n) &\leq card(\text{ shattered subsets of } \{x_1, \ldots, x_n\}) \\ &\leq card(\text{ subsets of size} \leq V) \\ &= \sum_{i=1}^{V} \binom{n}{i} \\ &\leq \left(\frac{n}{V}\right)^V \sum_{i=1}^{V} \binom{n}{i} \left(\frac{V}{n}\right)^i \\ &= \left(\frac{n}{V}\right)^V \left(1 + \frac{V}{n}\right)^V \\ &\leq \left(\frac{en}{V}\right)^V . \end{aligned} \tag{12}$$

□

Lemma 9 *Assume $\mathcal{H}$ is the half spase in $\mathbb{R}^d$, namely*

$$\mathcal{H} = \{x \to sign(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\},$$

then VC dimension of $\mathcal{H}$ equal to $d + 1$

Proof. We prove this lemma by proving two claims:

1. $VC(\mathcal{H}) \geq d + 1$
   Let $(x_1, \ldots, x_{d+1}) = (e_1, \ldots, e_d, 0)$ and let $(y_1, \ldots, y_{d+1}) \in \{\pm 1\}^{d+1}$ be the label.
   Then we construct a and b by
   $$a_i = y_i - y_{d+1}$$
   and
   $$b = y_{d+1}.$$
   Then we have
   $$sign(a^T e_i + b) = sign(y_i - y_{d+1} + y_{d+1}) = y_i$$
   and
   $$sign(a^T 0 + b) = sign(y_{d+1} = y_{d+1}.$$
   So this shows that $\mathcal{H}$ shatters $\{x_1, \ldots, x_{d+1}\}$, and thus $VC(\mathcal{H}) \geq d + 1$.

2. $VC(\mathcal{H}) < d + 2$

   By Radon theorem, any set of $d + 2$ points in $\mathbb{R}^d$ can by partitioned into two disjoint subsets whose convex hull have a non-empty intersection. So label one of the two partitions $+1$ and the other -1. No half-space can successfully label these points in the intersection.

□

6 Bound for $G_n(\mathcal{H})$

We know that

$$G_n(\mathcal{H}) := \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{H}}$$

Define

$$\begin{aligned}\overline{G_n(\mathcal{H})} &= \|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{H}}\\ &= \sup_{H\in\mathcal{H}} |\frac{1}{n}\sum_{i=1}^{n} g_i I(X_i\in H)|.\end{aligned}$$

And let

$$\mathcal{S} = \{s = (h(X_1),\ldots,h(X_n)) : h\in\mathcal{H}\}.$$

Then $\forall t$, we have

$$\begin{aligned}\mathbb{P}(\overline{G_n(\mathcal{H})} \geq t) &= \mathbb{P}(\sup_{s\in\mathcal{S}} |\frac{1}{n}\sum_{i=1}^{n} g_i s_i| \geq t)\\ &\leq \sum_{s\in\mathcal{S}} \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} g_i s_i| \geq t). \qquad (13)\end{aligned}$$

And it's obvious that, given $s = (s_1,\ldots,s_n)$,

$$\frac{1}{n}\sum_{i=1}^{n} g_i s_i \sim N(0, \frac{1}{n^2}\sum_{i=1}^{n} s_i^2)$$

So it follows that

$$\begin{aligned}\sum_{s\in\mathcal{S}} \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} g_i s_i| \geq t) &= 2\sum_{s\in\mathcal{S}} \mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} g_i s_i \geq t)\\ &\leq 2|\mathcal{S}|\mathbb{P}(Y\geq t)\\ &\leq 2|\mathcal{S}|e^{-\frac{nt^2}{2}} \qquad (14)\end{aligned}$$

where $Y \sim N(0, \frac{1}{n})$

Now we only need to set a bound for $|\mathcal{S}|$.

By Sauer's lemma (Appendix B), we can conclude that

$$|S| \leq \left(\frac{en}{V}\right)^V$$

where V is the VC dimension of $\mathcal{H}$.

Then by lemma 11 (Appendix B), we know that $V = d+1$, so

$$|S| \leq \left(\frac{en}{d+1}\right)^{d+1}$$

So

$$\mathbb{P}(\overline{G_n(\mathcal{H})} \geq t) \quad \leq \quad \max\{2\left(\frac{en}{d+1}\right)^{d+1} e^{-\frac{nt^2}{2}}, 1\}. \tag{15}$$

Let $v = d+1$ and $c = 2(\frac{e}{d+1})^{d+1}$, and we can simplify (15) by

$$\mathbb{P}(\overline{G_n(\mathcal{H})} \geq t) \leq \max\{ce^{v\log(n) - \frac{nt^2}{2}}, 1\}.$$

Then by simple calculation, we can get $t^* = \sqrt{\frac{4vlog(n)}{n}}$, such that $\forall t \geq t^*$,

$$v\log(n) - \frac{nt^2}{2} \leq -\frac{nt^2}{4}.$$

Finally, we can bound $G_n(\mathcal{H})$ by

$$\begin{aligned}
G_n(\mathcal{H}) &= \mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{H}} \\
&= \mathbb{E}\overline{G_n(\mathcal{H})} \\
&= \int_0^\infty \mathbb{P}(\overline{G_n(\mathcal{H})} \geq t)dt \\
&\leq \int_0^{t*} 1dt + c\int_{t^*}^\infty e^{-\frac{nt^2}{4}}dt && (16) \\
&\leq t^* + \frac{c}{\sqrt{n}}\int_0^\infty e^{-\frac{x^2}{2}}dx && (17) \\
&\leq \sqrt{\frac{4vlog(n)}{n}} + \frac{c}{\sqrt{n}}. && (18)
\end{aligned}$$

This is a very good bound, since when $n \to \infty$, $G_n(\mathcal{H}) \to 0$, in the speed of $\sqrt{\frac{\log(n)}{n}}$