EMPIRICAL MARGIN DISTRIBUTIONS AND BOUNDING THE GENERALIZATION ERROR OF COMBINED CLASSIFIERS

这篇paper的时间有点早，是“Beyond Dropout: Feature Map Distortion to Regularize Deep Neural Networks”这篇paper的主要参考文献，也是与集成学习相关。这篇paper理论比较多，可能很多地方理解的不到位，只是单纯拿来给自己看看的笔记，目前只看了其中一部分，剩余part没用到就没看了。

Background

从论文title可以看出来这篇paper的研究对象就是这种多个分类器模型（有两类，一类是bagging和boosting这类的集成学习，还有一类是神经网络）的泛化损失上界(bounds on generalization error)，泛化损失其实就是指训练集上的平均loss和全局数据的loss之间的差距，当然从分布上理解就是说看我们多个分类器的经验边缘分布和真实边缘分布之间的距离。

文章用到了Gaussian and empirical processes (comparison inequalities, symmetrization method, concentration inequalities)这些理论支撑。

Related work

- 1

之前的一些方法会去考虑，不同classfier的损失函数之间的熵或者VC维（下面这个图中公式的每一项代表一个classfier的结果，最后衡量这些量之间的关系），来去衡量最后的这个bound，但是对于神经网络这种情况来说，很多时候分类器的VC维会很大，甚至无穷。

$$\{\{(x,y) : yg(x) \leq 0\} : g \in \mathcal{G}\}$$

- 2

还有一些方法考虑使用分类器分类器边缘$Y\hat{g}(x)$的经验分布，最后得到的bound如下：

$$\inf_{\delta>0}\left[n^{-1}\sum_{j=1}^{n} I_{\{Y_j\hat{g}(X_j)\leq\delta\}} + C(\mathcal{G})\phi(\delta)\frac{C(\mathcal{H})}{\sqrt{n}}\right],$$

对于这个bound的使用上，根据观察，利用boost这种集成学习方法正确分类的样本往往$Y\hat{g}(x)$比较大，这样δ就可以在确保前面一项不升高的情况下变得比较大，而后面一项种那个$\phi(\delta)$这一项是与δ反比的，这样如果能选一个合适的δ，整个bound就可以降低了。这种方法有理有据效果不错，但是有学者提到它仍然没有实现最好的衡量bound的方法。

- 3

其它的还有考虑最优化里面凸包等等的一些定理的运用。

Introduction

这篇文章提出的方法一方面解释了之前的一些关于泛化上界的方法，用到了Gaussian和empirical processes的一些理论，最终是以一个margin loss的形式提出一个新的generalization bound，容易使用。

论文提到，最终提出的bound和之前的神经网络L1范数这种形式的bound形式很类似，有很大意义。

论文也给出了经验margin distribution以概率一收敛到真实的margin distribution的条件。

Bounding the generalization error of convex combinations of classifiers

这个是论文中第五节，讲了文中提出的bound在classification上的应用（并且是一般的multi class，Rademacher复杂度定义本身是描述的是二分类问题）。

- 1

先给出一些符号，有M个类，其余符号如下：

> Consider a class $\tilde{\mathcal{F}}$ of functions from $\tilde{S} := S \times \mathcal{Y}$ into $\mathbb{R}$. A function $f \in \tilde{\mathcal{F}}$ predicts a label $y \in \mathcal{Y}$ for an example $x \in S$ iff
>
> $$f(x, y) > \max_{y' \neq y} f(x, y').$$
>
> The margin of a labeled example (x, y) is defined as
>
> $$m_f(x, y) := f(x, y) - \max_{y' \neq y} f(x, y'),$$
>
> so f misclassifies the labeled example (x, y) iff $m_f(x, y) \leq 0$. Let
>
> $$\mathcal{F} := \{f(\cdot, y) : y \in \mathcal{Y}, f \in \tilde{\mathcal{F}}\}.$$

- 2

基于第二节里面的证明，文章先给出了一个定理：

> THEOREM 11. *For all $t > 0$,*
>
> $$\mathbb{P}\Big\{\exists f \in \tilde{\mathcal{F}} : P\{m_f \leq 0\} > \inf_{\delta \in (0,1]} \Big[P_n\{m_f \leq \delta\} + \frac{8M(2M-1)}{\delta} R_n(\mathcal{F}) + \Big(\frac{\log \log_2(2\delta^{-1})}{n}\Big)^{1/2}\Big] + \frac{t}{\sqrt{n}}\Big\} \leq 2\exp\{-2t^2\}.$$

这个定理用这种形式看还是优点难看懂的，和我们想象中的upper bound不大一样，不妨粗略处理一下看看，这里我自己想的不知道对不对，首先看左边概率里面不等式左边是真实地全局loss概率，不等式右边，右边取inf，那么第一项肯定是$\delta = 0$时候最小，取0后这不就是平均loss地概率吗，第二项看了后面就知道实际上$R_n(\hat{f}) <= M(2M-1) * R_n(f)$（前面那个不是f hat 上面是~，一下子latex没打出来），第三项取inf不就变成0了吗。

先证明一下，证明之前首先要给出下面这个引理：

For a class of functions $\mathcal{H}$, we will denote by

$$\mathcal{H}^{(l)} = \{\max(h_1, \dots, h_l) : h_1, \dots, h_l \in \mathcal{H}\}.$$

LEMMA 2. *The following bound holds:*

$$\mathbb{E}\|\sum_{i=1}^n \varepsilon_i \delta_{X_i}\|_{\mathcal{H}^{(l)}} \le 2l\mathbb{E}\|\sum_{i=1}^n \varepsilon_i \delta_{X_i}\|_{\mathcal{H}}.$$

证明这个引理，首先要证明下面这个公式：

Proof. Let $x^+ := x \vee 0$. Obviously $x \mapsto x^+$ is a nondecreasing convex function such that $(a+b)^+ \le a^+ + b^+$. We will first prove that

$$\mathbb{E}(\sup_{\mathcal{H}^{(l)}} \sum_{i=1}^n \varepsilon_i h(X_i))^+ \le l\mathbb{E}(\sup_{\mathcal{H}} \sum_{i=1}^n \varepsilon_i h(X_i))^+. \tag{5.1}$$

这里不是很难，分为两步，第一步比较简单，用一下简单的不等式拆分就可以了，这里注意一下，为什么可以做第一步不等式操作呢，文中没说，应该是默认了F1，F2都是属于F的：

Let us consider classes of functions $\mathcal{F}_1$, $\mathcal{F}_2$ and

$$\mathcal{F} = \{\max(f_1, f_2) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}.$$

Since

$$\max(f_1, f_2) = \frac{1}{2}\left((f_1 + f_2) + |f_1 - f_2|\right),$$

we have

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{\mathcal{F}} \sum_{i=1}^n \varepsilon_i f(X_i)\Big)^+ \\
&\le \mathbb{E}\Big(\sup_{\mathcal{F}_1,\mathcal{F}_2} \sum_{i=1}^n \varepsilon_i \frac{1}{2}(f_1(X_i) + f_2(X_i)) + \sup_{\mathcal{F}_1,\mathcal{F}_2} \sum_{i=1}^n \varepsilon_i \frac{1}{2}|f_1(X_i) - f_2(X_i)|\Big)^+ \\
&\le \frac{1}{2}\mathbb{E}\Big(\sup_{\mathcal{F}_1,\mathcal{F}_2} \sum_{i=1}^n \varepsilon_i (f_1(X_i) + f_2(X_i))\Big)^+ + \frac{1}{2}\mathbb{E}\Big(\sup_{\mathcal{F}_1,\mathcal{F}_2} \sum_{i=1}^n \varepsilon_i |f_1(X_i) - f_2(X_i)|\Big)^+ \\
&\le \frac{1}{2}\mathbb{E}\Big(\sup_{\mathcal{F}_1} \sum_{i=1}^n \varepsilon_i f_1(X_i)\Big)^+ + \frac{1}{2}\mathbb{E}\Big(\sup_{\mathcal{F}_2} \sum_{i=1}^n \varepsilon_i f_2(X_i)\Big)^+ \\
&+ \frac{1}{2}\mathbb{E}\Big(\sup_{\mathcal{F}_1,\mathcal{F}_2} \sum_{i=1}^n \varepsilon_i |f_1(X_i) - f_2(X_i)|\Big)^+.
\end{aligned}$$

现在不等式右边有三项其中第三项有绝对值，我们进一步化简去绝对值，来到第二步，第二步需要用前面证明得到的一个定理：

The proof of Theorem 4.12 in [24] contains the following statement. If T is a bounded subset of $\mathbb{R}^n$, functions φ_i, $i = 1, \dots, n$ are contractions such that $\varphi_i(0) = 0$ and a function $G : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing then

$$\mathbb{E}G\Big(\sup_{t\in T} \sum_{i=1}^n \varepsilon_i \varphi_i(t_i)\Big) \le \mathbb{E}G\Big(\sup_{t\in T} \sum_{i=1}^n \varepsilon_i t_i\Big).$$

用了这个定理以后，$(\cdot)^+$这个函数自然是非递减的凸函数，取绝对值也满足在0点为0，替换一下，在简单处理后就可以去掉第三个项含绝对值得，得到一个比较简洁的式子：

If we take $G(x) = x^+$, $\varphi_i(x) = |x|$ and $T = \{(f_1(X_i) - f_2(X_i))_{i=1}^n : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ we get (first conditionally on $(X_i)_{i=1}^n$ and then taking expectations)

$$\mathbb{E}\Big(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i|f_1(X_i) - f_2(X_i)|\Big)^+ \le \mathbb{E}\Big(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i(f_1(X_i) - f_2(X_i))\Big)^+$$
$$\le \mathbb{E}\Big(\sup_{\mathcal{F}_1}\sum_{i=1}^n \varepsilon_i f_1(X_i)\Big)^+ + \mathbb{E}\Big(\sup_{\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i f_2(X_i)\Big)^+,$$

where in the last inequality we used the fact that the sequence $(-\varepsilon_i)_{i=1}^n$ is equal in distribution to $(\varepsilon_i)_{i=1}^n$. Combining the bounds gives

$$\mathbb{E}\Big(\sup_{\mathcal{F}}\sum_{i=1}^n \varepsilon_i f(X_i)\Big)^+ \le \mathbb{E}\Big(\sup_{\mathcal{F}_1}\sum_{i=1}^n \varepsilon_i f_1(X_i)\Big)^+ + \mathbb{E}\Big(\sup_{\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i f_2(X_i)\Big)^+.$$

这个式子是针对$l = 2$的情形的，简单用一下归纳法就可以得到（5.1）这个式子了，再放一遍：

Proof. Let $x^+ := x \vee 0$. Obviously $x \mapsto x^+$ is a nondecreasing convex function such that $(a+b)^+ \le a^+ + b^+$. We will first prove that

$$\mathbb{E}(\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i))^+ \le l\mathbb{E}(\sup_{\mathcal{H}}\sum_{i=1}^n \varepsilon_i h(X_i))^+. \tag{5.1}$$

然后结合上我们随机变量ε取正负分布都一样，可以证明得到，这里前面没说，δ_x没理解错的话应该是狄拉克函数，不过这个第一步我还是没看懂，mark一下，我觉得是把δ_x拆开成了$h(x_i) - h(x_i)$，再不等式操作一下，但我不是很确定，不过后面几步利用上面刚证出的就很容易了：

$$\mathbb{E}\|\sum_{i=1}^n \varepsilon_i \delta_{X_i}\|_{\mathcal{H}^{(l)}} \le \mathbb{E}\Big(\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i)\Big)^+ + \mathbb{E}\Big(-\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i)\Big)^+$$
$$= 2\mathbb{E}\Big(\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i)\Big)^+ \le 2l\mathbb{E}\Big(\sup_{\mathcal{H}}\sum_{i=1}^n \varepsilon_i h(X_i)\Big)^+ \le 2l\mathbb{E}\|\sum_{i=1}^n \varepsilon_i \delta_{X_i}\|_{\mathcal{H}}.$$

到目前为止是把要用的引理证出来了，现在来看它正式的证明，下面几步操作都挺直观的：

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j m_f(X_j, Y_j)\Big| = \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j \sum_{y\in\mathcal{Y}} m_f(X_j, y) I_{\{Y_j = y\}}\Big|$$
$$\le \sum_{y\in\mathcal{Y}} \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j m_f(X_j, y) I_{\{Y_j = y\}}\Big|$$
$$\le \frac{1}{2}\sum_{y\in\mathcal{Y}} \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j m_f(X_j, y)(2I_{\{Y_j = y\}} - 1)\Big| + \frac{1}{2}\sum_{y\in\mathcal{Y}} \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j m_f(X_j, y)\Big|.$$

上面最后一步的目的其实是为了构造Rademacher变量，看下面的证明，下面这里后面几步我其实没看明白为什么：

Denote $\sigma_j(y) := 2I_{\{Y_j = y\}} - 1$. Given $\{(X_j, Y_j) : 1 \le j \le n\}$, the random variables $\{\varepsilon_j \sigma_j(y) : 1 \le j \le n\}$ are i.i.d. Rademacher. Hence, we have

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j m_f(X_j, y)(2I_{\{Y_j = y\}} - 1)\Big| = \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j \sigma_j(y) m_f(X_j, y)\Big|$$
$$= \mathbb{E}\mathbb{E}_\varepsilon \sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j \sigma_j(y) m_f(X_j, y)\Big| = \mathbb{E}\mathbb{E}_\varepsilon \sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j m_f(X_j, y)\Big|$$
$$= \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^n \varepsilon_j m_f(X_j, y)\Big|.$$

抛开上面这一步没看懂的证明，它目的是得到下面这个式子，主要就是把关于y求和拿到了均值外面来：

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,Y_j)\right| \le \sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)\right|.$$

关于上边不等式右边，考虑进一步缩放，把y也放进sup里面，我们最开始证出的引理，主要用在了最后一个不等号上，注意，这里面第一步是从$m_f(x,y)$定义来的：

Next, using Lemma 2, we get for all $y \in \mathcal{Y}$

$$\begin{aligned}\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)\right| &\le \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j f(X_j,y)\right| + \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j \max_{y'\neq y} f(X_j,y')\right| \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j f(X_j)\right| + \mathbb{E}\sup_{f\in\mathcal{F}^{(M-1)}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j f(X_j)\right| \\ &\le (2M-1)\mathbb{E}\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j f(X_j)\right|.\end{aligned}$$

结合上面两步，可以得到：

$$\begin{aligned}&\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,Y_j)\right| \le \sum_{y\in\mathcal{Y}}(2M-1)\mathbb{E}\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j f(X_j)\right| \\ &= M(2M-1)\mathbb{E}\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j f(X_j)\right|,\end{aligned}$$

到这里，论文说这个结论和Theorem的情况就一样了，因为前面没仔细看，这里不知道论文说的情况一样是怎么个一样法。

论文还提到后面都用二分类来说，不知道为什么又变成二分类了，不知道是不是为了简单方便。论文下面证了二分类下loss和f(x)的Rademacher复杂度是一样的，诶那上面应该就是说多分类下f(x)的Rademacher复杂度是lossf(x)的Rademacher复杂度的上界？不知道有没有理解错：

> In the rest of the paper, we assume that the set of labels is $\{-1,1\}$, so that $\tilde{S} := S \times \{-1,1\}$ and $\tilde{\mathcal{F}} := \{\tilde{f} : f \in \mathcal{F}\}$, where $\tilde{f}(x,y) := yf(x)$. P will denote the distribution of (X,Y), P_n the empirical distribution based on the observations $((X_1,Y_1),\ldots,(X_n,Y_n))$. Clearly, we have
>
> $$R_n(\tilde{\mathcal{F}}) = \mathbb{E}\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{i=1}^{n}\varepsilon_i Y_i f(X_i)\right| = \mathbb{E}\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_i f(X_i)\right|,$$
>
> where $\tilde{\varepsilon}_i := Y_i\varepsilon_i$. Since, for given $\{(X_i,Y_i)\}$, $\{\tilde{\varepsilon}_i\}$ and $\{\varepsilon_i\}$ have the same distribution, we get
>
> $$\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_i f(X_i)\right| = \mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|,$$
>
> which immediately implies $R_n(\tilde{\mathcal{F}}) = R_n(\mathcal{F})$.

后面还有些内容，一下子看不动了。。。。。