Advanced Topics in Artificial Intelligence: EE6180
Indian Institute of Technology Madras October 19, 2020
Instructor: Abhishek Sinha

Mid-Term Solution

• **Do not distribute these solutions outside the class**.

---

1. **(Upper bound for the Expected Norm)** (a) Using Cauchy-Schwartz inequality, we have for any $w \in \mathbb{R}^d$:

$$\max_{v:||v||_2 \leq 1} v^T w = ||w||_2$$

Hence,

$$\mathcal{G}(T^d(s)) = \mathbb{E}(\max_{v:||v||_0 \leq s, ||v||_2 \leq 1} v^T w) \leq \mathbb{E}(\max_{|S|=s} ||w_S||_2).$$

(b) From the standard L_2 norm concentration of Gaussian R.V.s as derived in the class, we have

$$\mathbb{P}(||w_S||_2 \geq \sqrt{s} + \delta) \leq e^{-\delta^2/2},$$

(c) From part (b), we conclude that $||w_S||_2 - \sqrt{s} \sim \mathsf{SG}(0,1)$. Moreover, there are $\binom{d}{s} \leq \left(\frac{ed}{s}\right)^s$. subsets of cardinality s. Hence, using the Maximal inequality for sub-Gaussians (Massart's lemma), we have

$$\mathcal{G}(T^d(s)) \leq \mathbb{E}(\max_{S:|S|=s} ||w_S||_2) \leq \sqrt{s} + \sqrt{2s \ln\left(\frac{ed}{s}\right)}.$$

2. **(Hedge with Many Good Experts)** Define the potential function

$$\Phi_t = \frac{1}{\eta} \ln\left(\sum_{i=1}^{N} \exp(-\eta L_t(i))\right).$$

Proceeding in the same way as shown in the class, we have

$$\hat{L}_T^{\mathtt{Hedge}} \leq \Phi_0 - \Phi_T + \eta T. \tag{1}$$

Next, we have

$$\Phi_0 = \frac{\ln N}{\eta},$$

and

$$\Phi_T \geq \frac{\ln N_L}{\eta} - L.$$

The result now follows upon substitution in Eqn. (1).

3. **(Arbitrarily Small Training Error)** Consider the `Hedge` algorithm. Let ϵ be the fraction of experts with cumulative loss at most L. From problem 2 we have

$$\hat{L}_T^{\texttt{Hedge}} \leq L + \frac{1}{\eta}\log(\frac{1}{\epsilon}) + \eta T.$$

Using the Boosting to Online Learning reduction as discussed in the class, from the above, we have

$$(\frac{1}{2} + \gamma)T \leq L + \frac{1}{\eta}\log(\frac{1}{\epsilon}) + \eta T.$$

Hence, for a fraction of $1-\epsilon$ examples, the classification accuracy over T rounds is lower bounded as

$$\frac{L}{T} \geq \frac{1}{2} + \gamma - \frac{1}{\eta T}\log(\frac{1}{\epsilon}) - \eta.$$

Choose $\eta = \sqrt{\frac{\log(1/\epsilon)}{T}}$. We find that if $\gamma > 2\sqrt{\frac{\log(1/\epsilon)}{T}}$, *i.e.,* $T \geq 4\frac{\log(1/\epsilon)}{\gamma^2}$, the classification accuracy is strictly larger than $\frac{1}{2}$. Hence, for at least $1-\epsilon$ fraction of the training sample, the majority vote classifier $H(x) = \texttt{sign}(\sum_{t=1}^{T}\frac{1}{T}h_t(x))$ does not make any mistake.

4. **(Group Testing Lower Bounds)** Let the random variable W denote the identity of the true vector containing k 1's and $n-k$ 0's. Assume that W is sampled uniformly at random from the set of all $\binom{n}{k}$ vectors. The vector W is group-tested with the query matrix Φ which results in a random m dimensional binary test-result $Y^m : Y_i \in 0, 1, \forall i$. Finally, let $\hat{W}$ is the inferred vector. The entire procedure may be represented as

$$W \to \Phi \to Y^m \to \hat{W}.$$

Using Fano's bound, we have

$$\mathbb{P}(W \neq \hat{W}) \geq 1 - \frac{I(W;\hat{W}) + 1}{\log\binom{n}{k}}.$$

Using Data-Processing inequality, $I(W;\hat{W}) \leq I(W;Y^m)$, the elementary fact $\binom{n}{k} \geq (\frac{n}{k})^k$, and the requirement that $\mathbb{P}(W \neq \hat{W}) \leq \epsilon$, we have the fundamental lower bound:

$$\epsilon \geq 1 - \frac{I(W;Y^m) + 1}{k\log(n/k)}. \tag{2}$$

(a) In the noiseless case, we have

$$I(W;Y^m) = H(Y^m) - \underbrace{H(Y^m|W)}_{=0} \leq m.$$

where we have used the fact that Y^m is a deterministic function of W and entropy of a random variable is bounded above by the logarithm of the cardinality of its support set. Hence, from Eqn. (2), we have

$$m \geq (1-\epsilon)k\log(n/k) - 1.$$

(b) Since given the true vector W, the testing results are i.i.d., we have

$$H(Y^m|W) = mH(Y_1|W) = mH(q).$$

Thus,

$$I(W;Y^m) = H(Y^m) - H(Y^m|W) \leq m - mh(q).$$

Hence, from Eqn. (2), we have

$$m \geq \frac{(1-\epsilon)k\log(n/k) - 1}{1-h(q)}.$$