CONCENTRATION OF MEASURE IN PROBABILITY AND HIGH-DIMENSIONAL STATISTICAL LEARNING – MASTER 2 – HOMEWORK 2

Problem

In this problem we consider a non-negative loss function $(z, h) \mapsto \ell(z, h)$ defined on $\mathcal{Z} \times \mathcal{H}$, where $\mathcal{H}$ is a hypothesis class. Given a learning algorithm $\mathcal{A}$ and a training sample $s = (z_1, \dots, z_n)$, we denote $h_s = \mathcal{A}(s)$ the hypothesis $h_s \in \mathcal{H}$ returned by $\mathcal{A}$ when applied on sample s. Recall that if $Z_1, \dots, Z_n$ are i.i.d. random variables on the sample set $\mathcal{X}$ with some fixed (but unknown) probability distribution $\mathbb{P}$, the empirical risk $\hat{\mathcal{R}}_S(h)$ of a hypothesis h on the random sample $S = (Z_1, \dots, Z_n)$ and its true risk $\mathcal{R}(h)$ are given by

$$\hat{\mathcal{R}}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, h), \quad \text{and} \quad \mathcal{R}(h) = \mathbb{E}_{Z \sim \mathbb{P}}\big[\ell(Z, h)\big] \ .$$

1. Part 1

We consider as learning algorithm $\mathcal{A}$ satisfying the following property: if two samples s and s' differ by a single point (e.g., if $s = (z_1, \dots, z_n)$ and $s' = (z_1', z_2, \dots, z_n)$) then

$$(\star) \qquad \forall z \in \mathcal{Z}, \big|\ell(z, h_s) - \ell(z, h_{s'})\big| \le \beta \ .$$

The goal is to control the true risk of h_S (with high probability) by its empirical risk.

(1) Consider $s = (z_1, \dots, z_n)$ and $s' = (z_1, \dots, z_{n-1}, z_n')$. Show that
$$\big|\mathcal{R}(h_s) - \mathcal{R}(h_{s'})\big| \le \beta \ .$$

(2) Show that if the loss ℓ is bounded by $M \ge 0$ then we also have
$$\big|\hat{\mathcal{R}}_s(h_s) - \hat{\mathcal{R}}_{s'}(h_{s'})\big| \le \beta + \frac{M}{n}.$$

(3) Consider the function $f : s = (z_1, \dots, z_n) \mapsto \mathcal{R}(h_s) - \hat{\mathcal{R}}_s(h_s)$. Show that if the loss is bounded by M then
$$\mathbb{P}\Big(f(S) \ge \epsilon + \mathbb{E}_S\big[f(S)\big]\Big) \le \exp\left(-\frac{2n\epsilon^2}{(2n\beta + M)^2}\right) .$$

(4) Here the goal is to prove the bound $\Big|\mathbb{E}_S\big[f(S)\big]\Big| \le \beta$.

  (a) Show that $\mathbb{E}_S\big[\mathcal{R}(h_S)\big] = \mathbb{E}_{(S,Z)\sim\mathbb{P}^{n+1}}\big[\ell(Z, h_S)\big]$.

  (b) Show that $\mathbb{E}_S\big[\hat{\mathcal{R}}_S(h_S)\big] = \mathbb{E}_{S\sim\mathbb{P}^n}\big[\ell(Z_1, h_S)\big]$.

Date: November 2020.

(c) Given a sample $s = (z_1, \dots, z_n)$ and $z \in \mathcal{Z}$ define $s' = g(s, z) = (z, z_2, \dots, z_n)$. Show that $\mathbb{E}_{(S,Z)\sim\mathbb{P}^{n+1}}\big[\ell(Z, h_{S'})\big] = \mathbb{E}_S\big[\hat{\mathcal{R}}_S(h_S)\big]$.

(d) Show that $\Big|\mathbb{E}_S\big[f(S)\big]\Big| \leq \beta$.

(5) Given a fixed risk $0 < \delta < 1$ (for example $\delta = 5\%$), deduce from the previous questions an upper bound U_n on $\mathcal{R}(h_S)$ such that $\mathbb{P}\big(\mathcal{R}(h_S) \leq U_n\big) \geq 1 - \delta$.

2. Part 2

In this section we exhibit an algorithm that satisfies the Hypothesis ($\star$) used in Part 1. Consider $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$, $\mathcal{X}$ the unit Euclidean ball of $\mathbb{R}^d$, $\mathcal{Y} = \{-1, +1\}$, and let the hypothesis class $\mathcal{H}$ be the set of linear forms $h : \mathbb{R}^d \to \mathbb{R}$ (identified with elements of $\mathbb{R}^d$) with Euclidean norm bounded by B. For $z = (x, y)$ and $h \in \mathcal{H}$ let $\ell(z, h) := \big(y - h(x)\big)^2$. Fix $\lambda > 0$. For a sample $s = (z_1, \dots, z_n) \in \mathcal{Z}^n$ define $F_s(h) = \hat{\mathcal{R}}_s(h) + \lambda\|h\|_2^2$. Let h_s be a minimizer of F_s and let $\mathcal{A}$ be the algorithm that outputs h_s when applied to sample s.

(1) Show that there exists $\sigma > 0$ such that for every $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and every $h_1, h_2 \in \mathcal{H}$:
$$\big|\ell(z, h_1) - \ell(z, h_2)\big| \leq \sigma\big|h_1(x) - h_2(x)\big| \;.$$

(2) Show that if $f : \mathbb{R}^d \to \mathbb{R}$ and $\lambda > 0$ are such that $x \mapsto f(x) - \lambda\|x\|^2$ is convex, and if u is a minimizer of f, then $\forall w \in \mathbb{R}^d, f(w) - f(u) \geq \lambda\|w - u\|^2$.

(3) Explain why F_s admits a minimizer on $\mathcal{H}$.

(4) Consider two samples s, s' differing but exactly one point (e.g. $z_n = (x_n, y_n)$ and $z'_n = (x'_n, y'_n)$). Let h be a minimizer of F_s and h' be a minimizer of $F_{s'}$.

(a) Show that $\big[F_s(h') - F_s(h)\big] - \big[F_{s'}(h') - F_{s'}(h)\big] \geq 2\lambda\|h' - h\|^2$.

(b) Deduce a lower bound on $\big[\hat{\mathcal{R}}_s(h') - \hat{\mathcal{R}}_s(h)\big] - \big[\hat{\mathcal{R}}_{s'}(h') - \hat{\mathcal{R}}_{s'}(h)\big]$.

(c) Show that
$$\big[\hat{\mathcal{R}}_s(h'_s) - \hat{\mathcal{R}}_s(h_s)\big] - \big[\hat{\mathcal{R}}_{s'}(h'_s) - \hat{\mathcal{R}}_{s'}(h_s)\big] \leq \frac{\sigma}{n}\Big(\big|h'(x_n) - h(x_n)\big| + \big|h'(x'_n) - h(x'_n)\big|\Big) \;.$$

(d) Deduce that $2\lambda\|h' - h\|^2 \leq \dfrac{2\sigma}{n}\|h' - h\|$.

(e) Prove that there exists β such that: for each $z \in \mathcal{Z}$,
$$\big|\ell(z, h') - \ell(z, h)\big| \leq \beta \;.$$

(5) How does the value of β obtained in the previous question depend on n ? Combined with the main conclusion of Part 1, how does the upper bound U_n depend on n ? Is it expected ?

(6) Can the result be adapted when $\mathcal{H}$ is the set of all linear forms (i.e. $B = \infty$) ?

Exercise 1

We denote by $|\cdot|$ the Euclidean norm on $\mathbb{R}^n$, and by B_n the corresponding unit ball. Recall that if $G \sim N(0, \mathrm{Id}_n)$, then for every 1-Lipschitz function $f : (\mathbb{R}^n, |\cdot|) \to \mathbb{R}$ and $t \geq 0$,

$$(\bigstar) \qquad \mathbb{P}\Big(f(G) \geq \mathbb{E}\big[f(G)\big] + t\Big) \leq \exp(-t^2/2).$$

We consider a Gaussian cloud $(G_i)_{1\leq i\leq N}$ where G_i are i.i.d. $N(0, \mathrm{Id}_n)$ random vectors and $N \gg n$. Let K be the convex hull of $\{G_i : 1 \leq i \leq N\}$.

(1) Show that $\mathbb{P}\big(|G| \geq \sqrt{n} + t\big) \leq \exp(-t^2/2)$. Deduce that if $N \geq e^n$, then
$$\mathbb{P}\Big(K \subset 3\sqrt{\log N} B_n\Big) \geq 1 - \frac{1}{N}.$$

(2) Show that if $(X_i)_{1\leq i\leq N}$ are i.i.d. $N(0,1)$ random variables, then
$$\mathbb{E}\Big[\max_{1\leq i\leq N} X_i\Big] \geq c\sqrt{\log N}$$
for some constant $c > 0$.

(3) Deduce from (2) and $(\bigstar)$ that for every $x \in S^{n-1}$,
$$\mathbb{P}\left(\sup_{y\in K}\langle x, y\rangle \leq \frac{c}{2}\sqrt{\log N}\right) \leq N^{-c^2/8}.$$

(4) Let $0 < \varepsilon < 1$. Show that if $P \subset S^{n-1}$ is an ε-separated set (with respect to the induced Euclidean distance), then
$$\varepsilon^n \operatorname{card}(P) \leq (1+\varepsilon)^n .$$
Conclude that S^{n-1} contains a ε-net (ε-dense) with cardinality $\leq (1 + 2/\varepsilon)^n \leq (3/\varepsilon)^n$.

(5) Choose a number $\alpha \geq e$ such that $\alpha^{c^2/8} > 36/c$. Show that if $N \geq \alpha^n$ then
$$\frac{c}{4}\sqrt{\log N} B_n \subset K \subset 3\sqrt{\log N} B_n$$
with large probability.
Hint: Use (3) for x in a $\frac{c}{12}$-net of S^{n-1}, and the union bound.

Exercise 2

An airline defines a suitcase (identified as a parallepiped) to be admissible if the sum of its dimensions (length+width+height) does not exceed 115 centimeters. Is it possible to hide a non-admissible suitcase inside an admissible suitcase?

Hint: Consider one of the notions introduced in Lecture 13.