High Dimensional Statistics

Spring 2025

Sharif University of Technology

Dr. Amir Najafi

Homework 3 | Uniform laws of large numbers and Metric entropy | Due: 1404/2/29

Problem 1: Glivenko–Cantelli Theorem *[7 points]*

Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables with common distribution function F, and define the empirical distribution function

$$F_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}\,.$$

Prove that

$$\sup_{x\in\mathbb{R}} \big|F_n(x) - F(x)\big| \;\longrightarrow\; 0 \quad \text{almost surely as } n \to \infty.$$

Problem 2: Failure of Glivenko–Cantelli *[8 points]*

Consider the class $\mathcal{S}$ of all subsets $S \subset [0,1]$ having a finite number of elements. Prove that the (empirical) Rademacher complexity of $\mathcal{F}_{\mathcal{S}}$ satisfies the lower bound

$$\mathcal{R}_n\big(\mathcal{F}_{\mathcal{S}}\big) = \mathbb{E}_{X,\varepsilon}\left[\sup_{S\in\mathcal{S}} \Big|\tfrac{1}{n}\sum_{i=1}^{n} \varepsilon_i\, \mathbb{1}_S(X_i)\Big|\right] \;\ge\; \tfrac{1}{2}.$$

Problem 3: VC Dimension of Closed and Convex Sets *[7 points]*

Let $\mathcal{C}^d_{\mathrm{cc}}$ be the class of all closed and convex subsets of $\mathbb{R}^d$. Show that $\mathcal{C}^d_{\mathrm{cc}}$ does *not* have finite VC dimension.

Problem 4: Generalization Bound Based on Covering Numbers *[8 points]*

Let H be a family of functions mapping $\mathcal{X}$ to a subset of real numbers $\mathcal{Y} \subseteq \mathbb{R}$. For any $\epsilon > 0$, the *covering number* $\mathcal{N}(H,\epsilon)$ of H for the $\|\cdot\|_\infty$ norm is the minimal $k \in \mathbb{N}$ such that H can be covered by k balls of radius ϵ, i.e. there exist $h_1, \ldots, h_k \in H$ with

$$\forall\, h \in H \;\; \exists\, i \le k :\; \|h - h_i\|_\infty = \sup_{x\in\mathcal{X}} |h(x) - h_i(x)| \le \epsilon.$$

In particular, if H is compact then $\mathcal{N}(H,\epsilon) < \infty$ for all $\epsilon > 0$.
Covering numbers measure the complexity of H. In this problem we prove a generalization

bound for the squared loss. Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$, and define for $h \in H$:

$$R(h) \;=\; \mathbb{E}_{(x,y)\sim D}\big[(h(x)-y)^2\big], \qquad \widehat{R}_S(h) \;=\; \frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2,$$

where $S = ((x_1,y_1),\ldots,(x_m,y_m)) \sim D^m$. Assume H is uniformly bounded: there exists $M > 0$ such that $|h(x) - y| \le M$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and all $h \in H$. Then one shows

$$\Pr_{S\sim D^m}\Big[\sup_{h\in H}\big|R(h)-\widehat{R}_S(h)\big| \;\ge\; \epsilon\Big] \;\le\; \mathcal{N}\big(H, \tfrac{\epsilon}{8M}\big)\, 2\exp\Big(-\tfrac{m\epsilon^2}{2M^4}\Big). \tag{1}$$

(a) Let

$$L_S(h) \;=\; R(h) - \widehat{R}_S(h).$$

Show that for all $h_1, h_2 \in H$ and any sample S,

$$\big|L_S(h_1) - L_S(h_2)\big| \;\le\; 4M\,\|h_1 - h_2\|_\infty.$$

(b) Suppose H can be covered by k subsets $B_1,\ldots,B_k$, i.e. $H = B_1 \cup \cdots \cup B_k$. Prove that

$$\Pr_{S\sim D^m}\Big[\sup_{h\in H}|L_S(h)| \ge \epsilon\Big] \;\le\; \sum_{i=1}^{k}\Pr_{S\sim D^m}\Big[\sup_{h\in B_i}|L_S(h)| \ge \epsilon\Big].$$

(c) Finally, let $k = \mathcal{N}(H, \frac{\epsilon}{8M})$ and choose $B_1,\ldots,B_k$ to be balls of radius $\frac{\epsilon}{8M}$ covering H. Using part (1), show for each $i = 1,\ldots,k$,

$$\Pr_{S\sim D^m}\Big[\sup_{h\in B_i}|L_S(h)| \ge \epsilon\Big] \;\le\; \Pr_{S\sim D^m}\big[\,|L_S(h_i)| \ge \tfrac{\epsilon}{2}\big].$$

Then apply Hoeffding's inequality to bound each term by $2\exp(-m\epsilon^2/(2M^4))$ and conclude (1).

Problem 5: Rademacher Control of Prediction Errors *[8 points]*

We are given a set of independent, identically distributed (i.i.d.) samples (X_i, Y_i) for $i = 1,\ldots,n$, drawn from some unknown probability distribution P_0 over a measurable space $\mathcal{X} \times \mathcal{Y}$. Based on this training data, a data-driven prediction set $\widehat{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ is constructed. That is, for each input $x \in \mathcal{X}$, the prediction set $\widehat{C}(x)$ is a subset of $\mathcal{Y}$ that we expect to contain the true label Y with high probability.
Additionally, fix a finite collection of subsets $\mathcal{G} = \{G_1,\ldots,G_m\}$ where each $G_j \subseteq \mathcal{X}$.
Define a function class $\mathcal{F}$ consisting of functions $f_{\boldsymbol{\beta}} : \mathcal{X} \to \{-1,0,1\}$ of the form:

$$f_{\boldsymbol{\beta}}(x) = \sum_{j=1}^{m}\beta_j\,\mathbb{1}_{\{x\in G_j\}},$$

where the coefficients $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_m)$ satisfy $\|\boldsymbol{\beta}\|_\infty \le 1$.

Next, for each $f \in \mathcal{F}$, define the random variable

$$Z_i(f) = f(X_i)\left(\mathbb{1}\{Y_i \in \widehat{C}(X_i)\} - (1-\alpha)\right),$$

where $\alpha \in (0,1)$ is a fixed constant. Note that $|Z_i(f)| \le 1$ for all i and f.
Our target quantity of interest is the worst-case bias over $\mathcal{F}$:

$$\Delta = \sup_{f \in \mathcal{F}} |P_0[Z_i(f)]|.$$

We also define its empirical analogue:

$$\widehat{\Delta} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(f) \right|.$$

Finally, let $\mathfrak{R}_n(\mathcal{F})$ denote the empirical Rademacher complexity of $\mathcal{F}$.

(a) Prove that for any $\delta \in (0,1)$, with probability at least $1-\delta$:

$$\Delta \le \widehat{\Delta} + 2\,\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

(b) Show that

$$\mathfrak{R}_n(\mathcal{F}) = \frac{1}{n} \sum_{j=1}^{m} \sqrt{n\,P_0(X \in G_j)} \quad \text{and} \quad \mathfrak{R}_n(\mathcal{F}) \le \sqrt{\frac{m}{n}}.$$

(Hint: Apply Cauchy–Schwarz and use that $\sum_j P_0(X \in G_j) \le 1$.)

(c) Combine your answers to parts (a) and (b) to deduce that, with probability at least $1-\delta$:

$$\Delta \le \hat{\Delta} + 2\sqrt{\frac{m}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

Problem 6: Failure of total boundedness *[7 points]*

Let $C([0,1], b)$ denote the class of all convex functions f defined on the unit interval such that $\|f\|_\infty \le b$. Show that $C([0,1], b)$ is *not* totally bounded in the sup-norm. *(Hint: Try to construct an infinite collection of functions $\{f^j\}_{j=1}^{\infty}$ such that $\|f^j - f^k\|_\infty \ge 1/2$ for all $j \ne k$.)*

Problem 7: Packing and covering *[7 points]*

Let $(\mathbb{T}, \rho)$ be a metric space. Prove the following inequalities between the packing and covering numbers:

(a) $M(2\delta; \mathbb{T}, \rho) \le N(\delta; \mathbb{T}, \rho)$, (b) $N(\delta; \mathbb{T}, \rho) \le M(\delta; \mathbb{T}, \rho)$.

Problem 8: From VC dimension to metric entropy *[8 points]*

In this exercise, we explore the connection between VC dimension and metric entropy. Given a set class $\mathcal{S}$ with finite VC dimension ν, we show that the function class $\mathcal{F}_{\mathcal{S}} := \{\mathbb{1}_S : S \in \mathcal{S}\}$ of indicator functions has metric entropy at most

$$N(\delta; \mathcal{F}_{\mathcal{S}}, L^1(\mathbb{P})) \leq K(\nu) \left(\frac{3}{\delta}\right)^{2\nu}, \tag{2}$$

for a constant $K(\nu)$.
Let $\{\mathbb{1}_{S_1}, \ldots, \mathbb{1}_{S_N}\}$ be a maximal δ-packing in the $L^1(\mathbb{P})$-norm, so that

$$\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|_1 = \mathbb{E}[|\mathbb{1}_{S_i}(X) - \mathbb{1}_{S_j}(X)|] > \delta \quad \text{for all} \quad i \neq j.$$

By Exercise 5.2, this N is an upper bound on the δ-covering number.

(a) Suppose that we generate n samples X_i, $i = 1, \ldots, n$, drawn i.i.d. from $\mathbb{P}$. Show that the probability that every set S_i picks out a different subset of $\{X_1, \ldots, X_n\}$ is at least

$$1 - \binom{N}{2}(1-\delta)^n.$$

(b) Using part (a), show that for $N \geq 2$ and $n = \frac{3 \log N}{\delta}$, there exists a set of n points from which $\mathcal{S}$ picks out at least N subsets, and conclude that

$$N \leq \left(\frac{3 \log N}{\delta}\right)^{\nu}.$$

(c) Use part (b) to show that the bound (2) holds with $K(\nu) := (2\nu)^{2\nu-1}$.

Problem 9: Concentration of Gaussian suprema *[7 points]*

Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean Gaussian process, and define $Z = \sup_{\theta \in \mathbb{T}} X_\theta$. Prove that

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \delta) \leq 2e^{-\frac{\delta^2}{2\sigma^2}},$$

where $\sigma^2 := \sup_{\theta \in \mathbb{T}} \operatorname{var}(X_\theta)$ is the maximal variance of the process.

Problem 10: Dudley's Entropy Integral and Lipschitz Classes *[8 points]*

Let $X = [0,1]^d$, and fix $L > 0$. Consider the function class

$$\mathcal{H}_L \;=\; \big\{\, f : X \to [0,1] \colon \forall\, x, x' \in X,\; |f(x) - f(x')| \leq L\|x - x'\|_2 \big\}.$$

Given a sample $\{x_i\}_{i=1}^n \subset X$, define the empirical L_2-pseudometric

$$d_n(f,g) \;=\; \sqrt{\frac{1}{n}\sum_{i=1}^n \big(f(x_i) - g(x_i)\big)^2}.$$

(a) Show that for every $\varepsilon > 0$,

$$\log N\big(\mathcal{H}_L,\, d_n,\, \varepsilon\big) \;\le\; \Big(\frac{C_d\, L}{\varepsilon}\Big)^d,$$

where C_d depends only on the dimension d.

(b) Recall that the empirical Rademacher complexity of $\mathcal{H}_L$ is

$$\hat{\mathfrak{R}}_n(\mathcal{H}_L) \;=\; \frac{1}{n}\,\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{f\in\mathcal{H}_L} \sum_{i=1}^{n} \varepsilon_i\, f(x_i)\Big],$$

where $\{\varepsilon_i\}$ are i.i.d. Rademacher signs. Use Dudley's entropy integral bound

$$\hat{\mathfrak{R}}_n(\mathcal{H}_L) \;\le\; \frac{12}{\sqrt{n}} \int_0^{\operatorname{diam}(\mathcal{H}_L)} \sqrt{\log N(\mathcal{H}_L,\, d_n,\, \varepsilon)}\, d\varepsilon$$

together with part 1 to prove

$$\hat{\mathfrak{R}}_n(\mathcal{H}_L) \;\le\; C'_d\, L\, n^{-1/d},$$

for some constant C'_d depending only on d.

(c) Combine the bound from part 2 with symmetrization and concentration to show that with probability at least $1-\delta$ over the draw of $\{x_i\}$,

$$\sup_{f\in\mathcal{H}_L} \Big|\, \mathbb{E}_{x\sim P}[f(x)] \;-\; \frac{1}{n}\sum_{i=1}^{n} f(x_i) \Big| \;\le\; C''_d \left(L\, n^{-1/d} \;+\; \sqrt{\tfrac{\ln(1/\delta)}{n}} \right).$$

Reading Material

Read the paper *Uniform convergence may be unable to explain generalization in deep learning* and answer the following problems.

Note: For simplicity and brevity, we have omitted the full definitions and notations from the paper.

Problem 11: Paper Summarization *[9 points]*

In this problem, you should write a clear and concise overview of the paper in **no more than two pages**. Please ensure that your summary:

- **Highlights the paper's main goals and contributions.** What are the authors aiming to achieve, and why is it important?
- **Describes the key technical ideas and methods.** How do the authors approach the problem? What techniques or frameworks do they employ?
- **Discusses the significance and implications of the results.** Why are these findings impactful, and how might they influence future research or applications?

Above all, your summary should be well-structured, straightforward, and demonstrate a thorough understanding of the paper's content.

Problem 12: Designing Your Own "Paradoxical" Dataset *[8 points]*

The paper's linear and sphere examples both admit a simple symmetry transform T on each training point that yields a "bad" dataset $S' = T(S)$ with the same distribution yet zero empirical accuracy.
Invent a third synthetic data distribution D (in $\mathbb{R}^d$), together with a simple overparameterized model and GD-style training, for which you can define a non-trivial mapping T satisfying all of the following:

1. T is its own inverse:
$$T\big(T(x)\big) \;=\; x \quad \forall x\,.$$
2. If the learned classifier h_S has low error on random test data drawn from D, then it misclassifies *all* points in the transformed dataset:
$$\forall (x,y) \in S, \quad L\big(h_S, T(x), y\big) = 1\,.$$
3. The transformed training set has the same distribution:
$$T(S) \;\sim\; D^m\,.$$

Precisely describe:

(a) The data distribution $D \subseteq \mathbb{R}^d$.

(b) Your overparameterized model architecture (e.g. linear, ReLU net, random features).

(c) The symmetry transform T.

Show formally that your choice of D, model, algorithm, and T satisfies the three properties above.

Problem 13: Failure of Uniform Convergence with Gaussian Noise *[8 points]*

Fix integers $m \geq 2$, $K \geq 1$ and $D \geq 1$. Let $u \in \mathbb{R}^K$ be any unit vector and let $\sigma > 0$. The data distribution $\mathcal{D}$ on $\mathbb{R}^{K+D} \times \{-1, +1\}$ is defined as follows: draw the label y uniformly from $\{-1, +1\}$; conditioned on y set
$$x = \big(x_1, x_2\big) \in \mathbb{R}^{K+D}, \qquad x_1 = 2y\,u, \qquad x_2 \sim \mathcal{N}\big(0, \sigma^2\, I_D\big) \;\text{ independent of } y.$$
The learning rule is identical to that in the paper:
1. Initialise weights $w = (w_1, w_2) \in \mathbb{R}^{K+D}$ at the origin.
2. Given a training set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^m$ drawn i.i.d. from $\mathcal{D}$, perform one full pass of gradient ascent on the linear score $y\, h_w(x) = y\langle w, x\rangle$ with learning rate 1. Equivalently
$$w_1 = 2m\,u, \qquad w_2 = \sum_{i=1}^{m} y^{(i)} x_2^{(i)}.$$

Denote by h_S the classifier returned on the sample S.

(a) Show that for any fixed $\sigma > 0$ and every $\varepsilon, \delta \in (0, \frac{1}{4})$ there exists a constant $c = c(\sigma, \varepsilon, \delta)$ such that if $D \geq c\,m$ then with probability at least $1 - \delta$ (over $S \sim \mathcal{D}^m$)

$$\mathrm{gen}_{0\text{-}1}(h_S) \;=\; \Big|\, \Pr_{(x,y)\sim\mathcal{D}}\big[h_S(x) \neq y\big] \;-\; \tfrac{1}{m}\sum_{i=1}^{m} \mathbf{1}\big[h_S(x^{(i)}) \neq y^{(i)}\big] \Big| \;\leq\; \varepsilon.$$

(Hint: follow the argument of Lemma 3.1 in the paper, but keep explicit dependence on σ.)

(b) Let $\mathcal{H}_\delta := \{h_S \mid S \sim \mathcal{D}^m,\, S \in \mathcal{S}_\delta\}$ for some measurable set $\mathcal{S}_\delta \subset (\mathbb{R}^{K+D} \times \{-1, +1\})^m$ with $\Pr[S \notin \mathcal{S}_\delta] \leq \delta$. Prove that under the same condition $D \geq c\,m$ we have

$$\varepsilon_{\mathrm{alg}}^{\mathrm{UC}}(m, \delta) \;:=\; \sup_{S\in\mathcal{S}_\delta} \sup_{h\in\mathcal{H}_\delta} \Big|\, \mathbb{E}_{\mathcal{D}}\big[\mathbf{1}[h(x) \neq y]\big] \;-\; \tfrac{1}{m} \sum_{(x,y)\in S} \mathbf{1}[h(x) \neq y] \,\Big| \;\geq\; 1 - \varepsilon.$$

That is, *even after restricting to the (algorithm-dependent) set of outputs with small test error, every two-sided uniform convergence bound remains nearly vacuous.*