MATHEMATICS OF MACHINE LEARNING Part II
Revision Sheet RDS/Lent 2024

The schedules for Mathematics of Machine Learning changed in 2023 to remove some material on the bounded differences inequality, but also changed in the following ways:

- IB Statistics became a formal prerequisite so linear regression can be used in examples (the relevant material is here: `https://www.statslab.cam.ac.uk/~rds37/teaching/machine_learning/Notation.pdf`).
- 'Bias–variance decomposition' was explicitly added.

From past exams then, the following questions should be accessible:

- 2023: All questions.
- 2022: Paper 1 31J, Paper 4 30J
- 2021: Paper 1 31J, Paper 2 31J
- 2020: Paper 2 30J, Paper 4 30J

I have added four more Tripos style questions below that you may wish to attempt. [Note these are not necessarily of precisely the standard difficulty of Tripos questions.]

1. What does it mean for a random variable $W \in \mathbb{R}$ to be *sub-Gaussian* with parameter $\sigma > 0$? State an upper bound on $\mathbb{P}(W - \mathbb{E}W > t)$ for $t > 0$.

   Show that if $W_1, \ldots, W_n$ are independent and sub-Gaussian with parameter σ, then $\sum_{i=1}^n W_i/n$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$.

   State Hoeffding's Lemma.

   Now suppose matrix $X \in [-1,1]^{n\times p}$ with $p \geq 2$ has independent rows with $\mathbb{E}(X_{ij}X_{ik}) = \Sigma_{jk}$ for all i, j, k where $\Sigma \in \mathbb{R}^{p\times p}$. Let $\hat{\Sigma} = X^T X/n$. Show that with probability at least $1 - 2p^{-2}$,
   $$\max_{j,k} |\hat{\Sigma}_{jk} - \Sigma_{jk}| \leq 2\sqrt{2\log(p)/n}.$$

2. Suppose we have input–output pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, 1\}$. Consider the empirical risk minimisation problem using hinge loss and hypothesis class
   $$\mathcal{H} = \{x \mapsto x^T\beta : \beta \in C \subseteq \mathbb{R}^p\},$$
   where C is a non-empty closed convex set. Write down the objective function $f : \mathbb{R}^p \to \mathbb{R}$ of the optimisation problem and briefly explain why it is convex.

   Now take $C = \{x \in \mathbb{R}^p : x_j \geq 0 \text{ for } j = 1, \ldots, p\}$. Write down the (sub)gradient descent procedure for minimising f over $\beta \in C$ giving explicit forms for any subgradients and projections used.

   Let $\hat{\beta} \in C$ be a minimiser f over C and suppose that $\max_{i=1,\ldots,n} \|x_i\|_2 \leq M$. Prove that the output $\bar{\beta}$ of your procedure with k iterations initialised at a $\beta_1 \in \mathbb{R}^p$ and implemented with a fixed step size η you should specify satisfies
   $$f(\bar{\beta}) - f(\hat{\beta}) \leq \frac{M\|\hat{\beta} - \beta_1\|_2}{\sqrt{k}}.$$

   [*You may use standard properties of convex functions and projections onto closed convex sets without proof.*]

3. Given a hypothesis class $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$ and i.i.d. input–output pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$, define the *Rademacher complexity* $\mathcal{R}_n(\mathcal{H})$.

   Now suppose

$$\mathcal{H} = \Big\{x \mapsto \sum_{j=1}^{d} \beta_j \phi_j(x) : \beta \in \mathbb{R}^d \text{ and } \sum_{j=1}^{d} \gamma_j^2 \beta_j^2 \leq \lambda^2\Big\}.$$

   where $\phi_j : \mathcal{X} \to \mathbb{R}$ and $\gamma_j > 0$ for $j = 1, \ldots, d$. Let $C^2 = \mathbb{E}\left(\sum_{j=1}^{d}\{\phi_j(X_1)/\gamma_j\}^2\right)$. Show that

$$\mathcal{R}_n(\mathcal{H}) \leq \frac{\lambda C}{\sqrt{n}}.$$

   Let R_ϕ and $\hat{R}_\phi$ be the risk and empirical risk respectively for logistic loss, and let h^* and $\hat{h}$ be the respective minimisers over $\mathcal{H}$ (so $\hat{h}$ is the empirical risk minimiser). Show that

$$\mathbb{E}R_\phi(\hat{h}) - R_\phi(h^*) \leq \frac{2\lambda C}{\log(2)\sqrt{n}}.$$

4. Let $\mathcal{F}$ be a family of functions $f : \mathcal{Z} \to \{a, b\}$ with $a \neq b$. Given $z_{1:n} \in \mathcal{Z}^n$, what is the *empirical Rademacher complexity* $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ of $\mathcal{H}$? What is meant by the *VC dimension* $\mathrm{VC}(\mathcal{F})$ of $\mathcal{F}$?

   Now suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \{-1, 1\}$ are i.i.d. input–output pairs and consider performing empirical risk minimisation with misclassification loss over a class of classifiers $\mathcal{H}$. Let R and $\hat{R}$ denote the risk and empirical risk respectively. State an upper bound of $\mathbb{E}\sup_{h\in\mathcal{H}}(R(h) - \hat{R}(h))$ in terms of the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ of a class $\mathcal{F}$ related to $\mathcal{H}$ in a way you should specify.

   Let $\mathcal{B}$ be a family of functions $\phi : \mathbb{R} \to \{-1, 1\}$ given by

$$\mathcal{B} = \{u \mapsto \mathrm{sgn}(u - a),\ u \mapsto \mathrm{sgn}(a - u) : a \in \mathbb{R}\}.$$

   Compute $\mathrm{VC}(\mathcal{B})$. Let $u_1, \ldots, u_n \in \mathbb{R}$ and state an upper bound on $|\mathcal{B}(u_{1:n})|$.

   Now for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p) \in \mathcal{B}^p$ define $\mathcal{H}_{\boldsymbol{\phi}}$ by

$$\mathcal{H}_{\boldsymbol{\phi}} = \{v \mapsto \mathrm{sgn}(\beta_1\phi_1(v_1) + \cdots + \beta_p\phi_p(v_p)) : \beta_1, \ldots, \beta_p \in \mathbb{R}\}.$$

   Fix $x_1, \ldots, x_n \in \mathbb{R}^p$, and derive an upper bound on $|\mathcal{H}_{\boldsymbol{\phi}}(x_{1:n})|$.

   Let $\mathcal{H} := \cup_{\boldsymbol{\phi}\in\mathcal{B}^p}\mathcal{H}_{\boldsymbol{\phi}}$ and show that

$$|\mathcal{H}(x_{1:n})| \leq (n + 1)^{3p}.$$

   Finally conclude that

$$\mathbb{E}\sup_{h\in\mathcal{H}}(R(h) - \hat{R}(h)) \leq 2\sqrt{\frac{6p\log(n + 1)}{n}}.$$