

The following Tripos-style questions are not necessarily of uniform difficulty (some are somewhat harder or longer than typical exam questions, and some perhaps shorter) but should hopefully be helpful for exam preparation, and you may wish to discuss them in your revision supervision.

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, \dots, W_n are independent and sub-Gaussian with parameter σ , then $\sum_{i=1}^n W_i/n$ is sub-Gaussian with parameter σ/\sqrt{n} .

State Hoeffding's Lemma.

Now suppose matrix $X \in [-1, 1]^{n \times p}$ with $p \geq 2$ has independent rows with $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}X_{ik}) = \Sigma_{jk}$ 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. What does it mean for a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ to have the *bounded differences property* with constants L_1, \dots, L_d ?

State the *bounded differences inequality*.

Define function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$f(x_1, \dots, x_d) = \frac{2}{d(d-1)} \sum_{j=2}^d \sum_{k=1}^{j-1} g(x_j, x_k)$$

where $g : \mathbb{R}^2 \rightarrow (-b, b)$ for $b > 0$ is such that $g(u, v) = g(v, u)$. Let X_1, \dots, X_d be i.i.d. random variables. Prove that

$$\mathbb{P}(|f(X_1, \dots, X_d) - \mathbb{E}f(X_1, \dots, X_d)| \geq t) \leq 2 \exp\left(-\frac{dt^2}{8b^2}\right).$$

3. Let \mathcal{F} be a family of functions $f : \mathcal{Z} \rightarrow \{a, b\}$ with $a \neq b$. Given $z_{1:n} \in \mathcal{Z}^n$, let $\mathcal{F}(z_{1:n}) = \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\}$. What is the *empirical Rademacher complexity* $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$? What is meant by the *VC dimension* $\text{VC}(\mathcal{F})$ of \mathcal{F} ?

Now suppose $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \{-1, 1\}$ 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 $\phi : \mathbb{R} \rightarrow \{-1, 1\}$ given by

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

Show that $\text{VC}(\mathcal{B}) = 2$. Let $u_1, \dots, u_n \in \mathbb{R}$ and state an upper bound on $|\mathcal{B}(u_{1:n})|$.

Now for $\phi = (\phi_1, \dots, \phi_p) \in \mathcal{B}^p$ define the hypothesis class \mathcal{H}_ϕ of functions $h : \mathbb{R}^p \rightarrow \{-1, 1\}$ by

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

Fix $x_1, \dots, x_n \in \mathbb{R}^p$, and show that $|\mathcal{H}_\phi(x_{1:n})| \leq (n+1)^p$. [You may use any result from the course provided you state it clearly.]

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

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

Finally conclude that

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

[You may assume any results in the course about sub-Gaussian random variables provided you state them clearly.]

4. Suppose we have input–output pairs $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, 1\}$. 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. Briefly explain why the objective function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ of the optimisation problem is convex.

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

Let $\hat{\beta} \in \mathbb{R}^p$ be a minimiser of f over C and suppose that $\max_{i=1, \dots, n} \|x_i\|_2 \leq M$. 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.]

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

Now suppose

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

where $\phi_j : \mathcal{X} \rightarrow \mathbb{R}$ and $\gamma_j > 0$ for $j = 1, \dots, d$ are fixed. 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}_ϕ 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

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

[You may use any result from the course provided you state it clearly.]