MATHEMATICS OF MACHINE LEARNING Part II
Example Sheet 3 (of 3) Lent 2025

1. Given a class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$, let $\mathcal{H} = \{\text{sgn} \circ f : f \in \mathcal{F}\}$. Let $D = (X_i, Y_i)_{i=1}^n$ be n i.i.d. input–output pairs taking values in $\mathcal{X} \times \{-1, 1\}$. Show that for any $\hat{h} = \text{sgn} \circ \hat{f} \in \mathcal{H}$ depending on D (e.g. the ERM over $\mathcal{H}$), and any $\rho > 0$, we have that the misclassification risk $R(\hat{h})$ satisfies
$$\mathbb{E}R(\hat{h}) \leq \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\hat{f}(X_i)Y_i \leq \rho\}}\right) + \frac{2}{\rho}\mathcal{R}_n(\mathcal{F}).$$
[*Hint: Construct an appropriate surrogate loss ϕ such that $\mathbb{1}_{(-\infty,0]} \leq \phi \leq \mathbb{1}_{(-\infty,\rho]}$, and ϕ has Lipschitz constant $1/\rho$.*]

2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex function and suppose $C \subseteq \mathbb{R}^d$ is a convex set. Suppose $x_1, x_2 \in C$ satisfy $f(x_1) = f(x_2) = \inf_{x \in C} f(x)$. Show that $x_1 = x_2$.

3. Show that for a closed convex set $C \subseteq \mathbb{R}^d$, $\pi \in C$ is a projection of $x \in \mathbb{R}^d$ onto C if
$$(x - \pi)^\top (z - \pi) \leq 0 \qquad \text{for all } z \in C.$$

4. Show that $\partial\|\beta\|_1 = \{b : \text{ for each } j, b_j \in [-1, 1] \text{ and } b_j = \text{sgn}(\beta_j) \text{ if } \beta_j \neq 0\}$.

5. Show that $x \in \mathbb{R}^d$ minimises $f : \mathbb{R}^d \to \mathbb{R}$ if and only if $0 \in \partial f(x)$.

6. This question derives the form of the projection onto an ℓ_1-norm constraint set.
   (a) Fix $x \in \mathbb{R}^p$ and $\gamma > 0$, and let $g(\beta) := \|\beta - x\|_2^2/2 + \gamma\|\beta\|_1$. Show that g is minimised over $\beta \in \mathbb{R}^p$ by
   $$\beta_j^* = \max(|x_j| - \gamma, 0)\text{sgn}(x_j).$$
   [*Hint: Use 4 and 5.*]
   (b) Argue that if β^* above has $\|\beta^*\|_1 = \lambda$, then β^* is the projection of x onto the set $C = \{z : \|z\|_1 \leq \lambda\}$ i.e. $\beta^* = \pi_C(x)$.

7. Consider a version of stochastic gradient descent for minimising
$$f(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(h_\beta(x_i), y_i)$$
(assumed here to be differentiable) over $\beta \in C \subseteq \mathbb{R}^p$ where C is closed and convex, and let $\hat{\beta}$ be the minimiser. We take $U_1, \ldots, U_{k-1}$ uniformly distributed on $\{1, \ldots, p\}$ and writing $\beta^{(s)} \in \mathbb{R}^p$ for the sth iterate we take
$$\tilde{g}_s = e_{U_s} \frac{\partial f}{\partial \beta_{U_s}}\bigg|_{\beta^{(s)}}.$$
Show that under the setup of Theorem 23 on the convergence of gradient descent (so in particular we assume $\sup_{\beta \in C} \|\nabla f(\beta)\|_2 \leq L$, the output $\bar{\beta}$ of the algorithm set out above, with a suitable step size $\eta > 0$ you should specify, satisfies
$$\mathbb{E}f(\bar{\beta}) - f(\hat{\beta}) \leq 2LR\sqrt{\frac{p}{k}}.$$

8. The following result shows that the theory for stochastic gradient descent can be used to obtain some forms of generalisation error bounds. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. input–output pairs. Consider empirical risk minimisation with logistic loss ϕ where $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_2 \le C\}$ and $\mathcal{H} = \{x \mapsto x^\top \beta : \|\beta\|_2 \le \lambda\}$. Let π denote projection onto $\{\beta : \|\beta\|_2 \le \lambda\}$. Let $\beta_1 \in \mathbb{R}^p$ be the 0 vector and define iteratively for $i = 1, \ldots, n-1$,

$$g_i = Y_i X_i \phi'(Y_i X_i^\top \beta_i),$$
$$\beta_{i+1} = \pi(\beta_i - \eta g_i).$$

[Note the β_i above are vectors.] Let $\bar{\beta} = \frac{1}{n}\sum_{i=1}^n \beta_i$ and set $\bar{h}(x) = x^\top \bar{\beta}$. Show that for some step size $\eta > 0$ you should specify,

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

[*Hint: Write the risk itself in the form* $\mathbb{E}\tilde{f}(\beta; U)$ *for some* U.]

9. Consider the Adaboost algorithm with base class $\mathcal{B}$ (satisfying that if $h \in \mathcal{B}$ then $-h \in \mathcal{B}$) and assume that at no iteration does any $h \in \mathcal{B}$ perfectly classify the data.

   (a) Show that

   $$\frac{\sum_{i=1}^n w_i^{(m+1)}}{\sum_{i=1}^n w_i^{(m)}} = 2\sqrt{\widehat{\text{err}}_m(1 - \widehat{\text{err}}_m)}$$

   where $\widehat{\text{err}}_m := \text{err}_m(\hat{h}_m)$.

   (b) Assume that for each m, $\widehat{\text{err}}_m \le 1/2 - \gamma$ for some $\gamma > 0$. Show that the empirical risk of the Adaboost output decreases exponentially fast with M:

   $$\frac{1}{n}\sum_{i=1}^n \exp(-Y_i \hat{f}_M(X_i)) = \prod_{m=1}^M 2\sqrt{\widehat{\text{err}}_m(1 - \widehat{\text{err}}_m)} \tag{1}$$
   $$\le \exp(-2\gamma^2 M).$$

   (c) Let

   $$\mathcal{H} = \left\{ \sum_{m=1}^M \beta_m h_m : \|\beta\|_1 \le 1,\ h_m \in \mathcal{B} \text{ for } m = 1, \ldots, M \right\}.$$

   Explain why for $x_{1:n} \in \mathcal{X}^n$,

   $$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \le \sqrt{\frac{2\text{VC}(\mathcal{B})\log(n+1)}{n}}.$$

   (d) Given input–output pairs $(X_i, Y_i)_{i=1}^n$ taking values in $\mathcal{X} \times \{-1, 1\}$, let $\hat{f}_M = \sum_{m=1}^M \hat{\beta}_m \hat{h}_m$ be the output of the Adaboost algorithm with base classifier class $\mathcal{B}$. With the assumption of (b) that $0 < \widehat{\text{err}}_m \le 1/2 - \gamma$ for some $1/2 > \gamma > \rho > 0$, show that

   $$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\hat{f}_M(X_i)Y_i \le \rho\|\hat{\beta}\|_1\}} \le \exp\{-2M(\gamma^2 - c\rho)\}, \quad \text{where} \quad c = \frac{1}{4}\log\left(\frac{1+2\gamma}{1-2\gamma}\right),$$

   and $\hat{\beta} := (\hat{\beta}_m)_{m=1}^M$. [*Hint: Use* $\mathbb{1}_{\{u \le b\}} \le \exp(b - u)$ *and* (1). *You may further use the fact that* $u \mapsto u^{1-\rho}(1-u)^{1+\rho}$ *is increasing for* $0 < u < 1/2 - \rho$.]

(e) Show that writing $\hat{h} := \mathrm{sgn} \circ \hat{f}_M$, we have that the misclassification risk $R(\hat{h})$ satisfies

$$\mathbb{E}R(\hat{h}) \leq \exp\{-2M(\gamma^2 - c\rho)\} + \frac{2}{\rho}\sqrt{\frac{2\mathrm{VC}(\mathcal{B})\log(n+1)}{n}}.$$

[*Hint: Recall Qu. 1.*]

10. Let $\mathcal{X} = \mathbb{R}^p$. Consider performing gradient boosting with base regression procedure the empirical risk minimiser over $\mathcal{H} = \{x \mapsto \mu + x_j\beta : j \in \{1, \ldots, p\},\ \mu, \beta \in \mathbb{R}\}$ with squared error loss. Consider the mth iteration. Show that $\hat{g}_m(x) = \hat{\mu}_{\hat{j}} + x_{\hat{j}}\hat{\beta}_{\hat{j}}$ where

$$\hat{\beta}_j := \frac{\sum_{i=1}^n \left(W_i - \bar{W}\right)\left(X_{ij} - \bar{X}_j\right)}{\sum_{i=1}^n \left(X_{ij} - \bar{X}_j\right)^2},$$
$$\hat{\mu}_j := \bar{W} - \hat{\beta}_j \bar{X}_j,$$

with $\bar{X}_j := \frac{1}{n}\sum_{i=1}^n X_{ij}$ and $\bar{W} := \frac{1}{n}\sum_{i=1}^n W_i$, and $\hat{j}$ maximises $|\hat{\rho}_j|$ over $j = 1, \ldots, p$ with

$$\hat{\rho}_j := \frac{\sum_{i=1}^n (W_i - \bar{W})(X_{ij} - \bar{X}_j)}{\left(\sum_{i=1}^n (W_i - \bar{W})^2\right)^{1/2}\left(\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2\right)^{1/2}}.$$

11. Consider the optimisation problem of performing a weighted empirical risk minimisation over the class of decision stumps. Specifically, suppose we have weights $w_1, \ldots, w_n > 0$ and a single predictor whose observations have been sorted as $X_1 < \cdots < X_n$. Show then that finding an ERM over

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

(i.e. minimising $\sum_{i=1}^n w_i \mathbb{1}_{\{h(X_i) \neq Y_i\}}$) may be performed in $O(n)$ computational operations.

12. In this question, we will study the Rademacher complexity of a simple neural network with a single hidden layer of m nodes, reLU activation function ψ, and additional ℓ_2-norm constraints on the parameters. Specifically, consider the set $\mathcal{H}$ of hypotheses of the form

$$h(x) = \sum_{j=1}^m \alpha_j \psi(\beta_j^\top x)$$

where $\alpha_j \in \mathbb{R}$ and $\beta_j \in \mathbb{R}^p$ for $j = 1, \ldots, m$, with the constraints $\|\alpha\|_2 \leq \lambda_\alpha$ and $\max_{j=1,\ldots,m} \|\beta_j\|_2 \leq \lambda_\beta$. Let $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_2 \leq C\}$ and take $x_{1:n} \in \mathcal{X}^n$.

(a) By considering $\mathcal{B} := \{x \mapsto \psi(b^\top x) : \|b\|_2 \leq \lambda_\beta\}$ and $\mathcal{B}' = \{x \mapsto b^\top x : \|b\|_2 \leq \lambda_\beta\}$, show that

$$\mathbb{E}\left(\sup_{b : \|b\|_2 \leq \lambda_\beta} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \psi(b^\top x_i)\right|\right) \leq \frac{2C\lambda_\beta}{\sqrt{n}},$$

where $\varepsilon_{1:n}$ are i.i.d. Rademacher random variables. [*Hint: Apply the contraction lemma.*]

(b) Let us introduce the set of vector-valued functions $\mathcal{G} := \{g := (g_1, \ldots, g_m) : g_j \in \mathcal{B} \text{ for } j = 1, \ldots, m\}$. Show that

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) = \lambda_\alpha \mathbb{E}\left(\sup_{g \in \mathcal{G}} \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\right\|_2\right).$$

(c) Finally show that

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \leq 2C\lambda_\alpha\lambda_\beta\sqrt{\frac{m}{n}}.$$