Part II

Mathematics of Machine Learning

Year

2023

2022

2021

2020

Paper 1, Section II

31J Mathematics of Machine Learning

(a) What does it mean for a set $C \subseteq \mathbb{R}^d$ to be *convex*?

(b) What does it mean for a function $f : C \to \mathbb{R}$ to be *strictly convex*? Show that any minimiser of f must be unique.

(c) Define the *projection* $\pi_C(x)$ of a point $x \in \mathbb{R}^d$ onto a closed convex set C. Briefly explain why this is unique. [Standard results about convex functions may be used without proof, and you need not show that $\pi_C(x)$ always exists.]

(d) Prove that $\pi \in C$ is the projection of x onto a closed convex set C if

$$(x - \pi)^T (z - \pi) \leqslant 0 \qquad \text{for all } z \in C.$$

(e) Let C be a closed convex set given by

$$C := \left\{ \begin{pmatrix} v \\ s \end{pmatrix} \in \mathbb{R}^p \times \mathbb{R} : \ \|v\|_2 \leqslant s \right\}.$$

Using part (d) or otherwise, show that if $(u, t) \in \mathbb{R}^p \times \mathbb{R}$ satisfy $\|u\|_2 \geqslant |t|$ then

$$\pi_C \left(\begin{pmatrix} u \\ t \end{pmatrix} \right) = \frac{1}{2} \left(1 + \frac{t}{\|u\|_2} \right) \begin{pmatrix} u \\ \|u\|_2 \end{pmatrix}.$$

What is $\pi_C \left(\begin{pmatrix} u \\ t \end{pmatrix} \right)$ when $\|u\|_2 \leqslant -t$?

(f) Let C be as in (e) and let $(X_i, Y_i) \in \mathbb{R}^{p+1} \times \mathbb{R}$ for $i = 1, \ldots, n$ be data formed of input–output pairs. Write down the projected gradient descent procedure for finding the empirical risk minimiser with squared error loss over the hypothesis class $\mathcal{H} = \{h : h(x) = \beta^T x, \text{ where } \beta \in C\}$, giving explicit forms for any gradients or projections involved.

Paper 2, Section II

31J Mathematics of Machine Learning

(a) Let $\mathcal{H}$ be a hypothesis class of functions $h : \mathcal{X} \to \{-1, 1\}$ with $|\mathcal{H}| > 2$ and $\mathcal{X} = \mathbb{R}^p$. Define the *shattering coefficient* $s(\mathcal{H}, n)$ and the *VC dimension* $\mathrm{VC}(\mathcal{H})$ of $\mathcal{H}$.

(b) Explain why if $\mathcal{H}_1, \mathcal{H}_2$ are hypothesis classes as above, then $s(\mathcal{H}_1 \cup \mathcal{H}_2, n) \leqslant s(\mathcal{H}_1, n) + s(\mathcal{H}_2, n)$.

Let us use the notation that, for a class $\mathcal{F}$ of functions $f : \mathbb{R}^p \to \mathbb{R}$, we write

$$\mathcal{H}_{\mathcal{F}} := \{h : h(x) = \mathrm{sgn} \circ f(x), \text{ where } f \in \mathcal{F}\}$$

for the class of functions derived through composition with the sgn function.

(c) Now let $\mathcal{F}_1 := \{f : f(x) = x^T\beta, \text{ where } \beta \in \mathbb{R}^p\}$. Stating any results from the course you need, show that

$$s(\mathcal{H}_{\mathcal{F}_1}, n) \leqslant (n+1)^p.$$

(d) Next for a class $\mathcal{G}$ of functions $g : \mathbb{R}^p \to \{-1, 1\}$, define for some fixed $m \in \mathbb{N}$,

$$\mathcal{F}_2 := \left\{ f : f(x) = \sum_{j=1}^{m} \alpha_j g_j(x), \text{ where } g_j \in \mathcal{G},\ \alpha \in \mathbb{R}^m \right\}.$$

Show that if $|\mathcal{G}| < \infty$,

$$s(\mathcal{H}_{\mathcal{F}_2}, n) \leqslant (n+1)^m |\mathcal{G}|^m.$$

Show furthermore that even if $|\mathcal{G}| = \infty$, we have

$$s(\mathcal{H}_{\mathcal{F}_2}, n) \leqslant (n+1)^m s(\mathcal{G}, n)^m.$$

[*Hint: Fix $x_{1:n} \in \mathcal{X}^n$ and consider $\mathcal{G}'$ with $|\mathcal{G}'| \leqslant s(\mathcal{G}, n)$ and $\mathcal{G}'(x_{1:n}) = \mathcal{G}(x_{1:n})$.*]

(e) Finally let $\mathcal{F}_3$ be the class of functions $f : \mathbb{R}^p \to \mathbb{R}$ given by a neural network with a single hidden layer of m nodes and activation function given by sgn. Show that

$$s(\mathcal{H}_{\mathcal{F}_3}, n) \leqslant (n+1)^{(p+1)m}.$$

Paper 4, Section II

30J Mathematics of Machine Learning

(a) Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R} \times \mathbb{R}$ be input–output pairs with $n \geqslant 4$. Describe the optimisation problem that a *regression tree algorithm* using a *squared error loss splitting criterion* would take to find the first split point.

(b) Assuming that the inputs are sorted so that $X_1 < \cdots < X_n$, show that the above may be solved in $O(n)$ computational operations.

(c) Now write down the squared error loss empirical risk minimiser $\hat{f}_m : \mathbb{R} \to \mathbb{R}$ over $\mathcal{F} : \{x \mapsto \alpha + x\beta \, : \alpha \in \mathbb{R}, \, \beta \in \mathbb{R}\}$, when trained only on data $(X_1, Y_1), \ldots, (X_m, Y_m)$ for $m \geqslant 2$. [You need not derive it.]

(d) Denote by $\hat{g}_m : \mathbb{R} \to \mathbb{R}$ the equivalent of $\hat{f}_m$ in part (c) when instead training only on $(X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)$ for $m \leqslant n - 2$. Show carefully how minimising

$$\sum_{i=1}^{m} (Y_i - \hat{f}_m(X_i))^2 + \sum_{i=m+1}^{n} (Y_i - \hat{g}_m(X_i))^2$$

over $m = 2, \ldots, n-2$ may be performed in $O(n)$ computations.

Paper 1, Section II

31J Mathematics of Machine Learning

(a) Let $\mathcal{F}$ be a family of functions $f : \mathcal{X} \to \{0, 1\}$ with $|\mathcal{F}| \geqslant 2$.

Define the *shattering coefficient* $s(\mathcal{F}, n)$ and the *VC dimension* $\mathrm{VC}(\mathcal{F})$ of $\mathcal{F}$.

State the *Sauer–Shelah lemma*.

(b) (i) Let

$$\mathcal{A}_1 = \left\{ \bigcup_{k=1}^{m} [a_k, b_k] : a_k, b_k \in \mathbb{R} \text{ for } k = 1, \ldots, m \right\}.$$

Show that $\mathcal{F}_1 := \{\mathbf{1}_A : A \in \mathcal{A}_1\}$ satisfies $\mathrm{VC}(\mathcal{F}_1) = m + 1$.

(ii) Let $\mathcal{F}_2$ be a class of functions from $\mathbb{R}^p$ to $\{0, 1\}$ given by

$$\mathcal{F}_2 := \{x \mapsto \mathbf{1}_{(0,\infty)}(\mu + x^T\beta) : \beta \in \mathbb{R}^p,\ \mu \in \mathbb{R}\}.$$

Stating any result from the course you need, give an upper bound on $\mathrm{VC}(\mathcal{F}_2)$.

(c) (i) Let $\mathcal{G}$ be a family of functions $g : \mathcal{Z} \to \{0, 1\}$ with $|\mathcal{G}| \geqslant 2$ and define $\mathcal{H}$ to be the set of functions $h : \mathcal{X} \times \mathcal{Z} \to \{0, 1\}$ for which $h(x, z) = f(x)g(z)$ for some $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Show that $s(\mathcal{H}, n) \leqslant s(\mathcal{F}, n)s(\mathcal{G}, n)$.

(ii) Now let $\mathcal{G}$ be a family of functions $g : \mathcal{X} \to \{0, 1\}$ with $|\mathcal{G}| \geqslant 2$ and define $\mathcal{H}$ to be the set of functions $h : \mathcal{X} \to \{0, 1\}$ for which $h(x) = f(x)g(x)$ for some $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Show that $s(\mathcal{H}, n) \leqslant s(\mathcal{F}, n)s(\mathcal{G}, n)$.

(d) (i) Let

$$\mathcal{A}_3 = \left\{ \prod_{j=1}^{p} \Big(\bigcup_{k=1}^{m} [a_{jk}, b_{jk}] \Big) : a_{jk}, b_{jk} \in \mathbb{R} \text{ for } j = 1, \ldots, p,\ k = 1, \ldots, m \right\}.$$

Show that $\mathcal{F}_3 := \{\mathbf{1}_A : A \in \mathcal{A}_3\}$ satisfies $s(\mathcal{F}_3, n) \leqslant (n + 1)^{(m+1)p}$.

(ii) For $m \geqslant 3$, let $\mathcal{A}_4$ be the set of all convex polygons in $\mathbb{R}^2$ with m sides, and set $\mathcal{F}_4 := \{\mathbf{1}_A : A \in \mathcal{A}_4\}$. Show that $s(\mathcal{F}_4, n) \leqslant (n + 1)^{3m}$.

Paper 2, Section II
31J Mathematics of Machine Learning

(a)What does it mean for a function $f : \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n \to \mathbb{R}$ to have the *bounded differences property* with constants $L_1, \ldots, L_n$?

State the *bounded differences inequality.*

(b) Let $\mathcal{X}$ and $\mathcal{Y}$ be input and output spaces respectively. Let H be a machine learning algorithm taking as its argument a dataset $D \in (\mathcal{X} \times \mathcal{Y})^n$ to output a hypothesis $H_D : \mathcal{X} \to \mathbb{R}$. For $D = (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, for all $i = 1, \ldots, n$ we write

$$D_i(x, y) := ((x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x, y), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)) .$$

Let $\ell : \mathbb{R} \times \mathcal{Y} \to [0, M]$ be a bounded loss function. Suppose H has the following property: there exists $\beta \geqslant 0$ such that for all $i = 1, \ldots, n$ and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\sup_{(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}} |\ell(H_{D_i(x,y)}(\tilde{x}), \tilde{y}) - \ell(H_D(\tilde{x}), \tilde{y})| \leqslant \beta.$$

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random input–output pair. Show that $F : (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}$ given by

$$F((x_1, y_1), \ldots, (x_n, y_n)) = \mathbb{E}\ell(H_D(X), Y) - \frac{1}{n} \sum_{i=1}^n \ell(H_D(x_i), y_i)$$

satisfies a bounded differences property with constants all equal to $2\beta + M/n$. [In the expectation above, the (x_i, y_i) are considered deterministic.]

(c) Now suppose $D = (X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ is a collection of i.i.d. input–output pairs independent of, and each having the same distribution as, (X, Y). Show that $\mathbb{E}F(D) \leqslant \beta$. [*Hint: Find an alternative expression for* $\mathbb{E}\ell(H_D(X), Y)$ *as a sum of expectations with the* i*th term involving* $H_{D_i(X,Y)}$.]

(d) Hence conclude that, given $0 < \delta \leqslant 1$,

$$\frac{1}{n} \sum_{i=1}^n \ell(H_D(X_i), Y_i) + \beta + (2n\beta + M)\sqrt{\frac{\log(1/\delta)}{2n}} \geqslant \mathbb{E}\ell(H_D(X), Y)$$

with probability at least $1 - \delta$.

Paper 4, Section II
30J Mathematics of Machine Learning

Throughout this question, you may assume that the optimum is achieved in any relevant optimisation problems, so for instance in part (a) you may assume $\hat{f}$ is well-defined.

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$ are i.i.d. input–output pairs. Let $\mathcal{B}$ be a set of classifiers $h : \mathcal{X} \to \{-1, 1\}$ such that $h \in \mathcal{B} \Rightarrow -h \in \mathcal{B}$.

(a) Write down the *Adaboost* algorithm using $\mathcal{B}$ as the base set of classifiers with tuning parameter M, which produces $\hat{f} : \mathcal{X} \to \mathbb{R}$ of the form $\hat{f} = \sum_{m=1}^{M} \hat{\beta}_m \hat{h}_m$ where $\hat{\beta}_m \geqslant 0$ and $\hat{h}_m \in \mathcal{B}$ for $m = 1, \ldots, M$. [You need not derive explicit expressions for $\hat{\beta}_m$ or $\hat{h}_m$.]

(b) For a set $S \subseteq \mathbb{R}^d$, what is meant by the *convex hull*, $\operatorname{conv} S$? What does it mean for a vector $v \in \mathbb{R}^d$ to be a *convex combination* of vectors $v_1, \ldots, v_m \in \mathbb{R}^d$? State a result relating convex hulls and convex combinations.

(c) Let ϕ denote the exponential loss. What is meant by the *ϕ-risk* $R_\phi(f)$ of $f : \mathcal{X} \to \mathbb{R}$? What is the corresponding *empirical ϕ-risk* $\hat{R}_\phi(f)$? Let $x_{1:n} \in \mathcal{X}^n$. What is meant by the *empirical Rademacher complexity* $\hat{\mathcal{R}}(\mathcal{B}(x_{1:n}))$?

(d) Consider a modification of the Adaboost algorithm where, if at any iteration $m \leqslant M$ we have $\sum_{k=1}^{m} \hat{\beta}_k > 1$, we terminate the algorithm and output $\hat{f} := \sum_{k=1}^{m-1} \hat{\beta}_k \hat{h}_k$, or the zero function if $m = 1$; otherwise we output $\hat{f} = \sum_{k=1}^{M} \hat{\beta}_k \hat{h}_k$ as usual. Let $r_{\mathcal{B}} = \sup_{x_{1:n} \in \mathcal{X}^n} \hat{\mathcal{R}}(\mathcal{B}(x_{1:n}))$. Show that

$$\mathbb{E} R_\phi(\hat{f}) \leqslant \mathbb{E} \hat{R}_\phi(\hat{f}) + 2 \exp(1) r_{\mathcal{B}}.$$

[*Hint: Introduce*

$$\mathcal{H} := \left\{ \sum_{m=1}^{M} \beta_m h_m : \sum_{m=1}^{M} \beta_m \leqslant 1,\ \beta_m \geqslant 0,\ h_m \in \mathcal{B} \text{ for } m = 1, \ldots, M \right\}.$$

You may use any results from the course without proof, but should state or name any result you use.]

Paper 1, Section II

31J Mathematics of Machine Learning

Let $\mathcal{H}$ be a family of functions $h : \mathcal{X} \to \{0, 1\}$ with $|\mathcal{H}| \geqslant 2$. Define the *shattering coefficient* $s(\mathcal{H}, n)$ and the *VC dimension* $\mathrm{VC}(\mathcal{H})$ of $\mathcal{H}$.

Briefly explain why if $\mathcal{H}' \subseteq \mathcal{H}$ and $|\mathcal{H}'| \geqslant 2$, then $\mathrm{VC}(\mathcal{H}') \leqslant \mathrm{VC}(\mathcal{H})$.

Prove that if $\mathcal{F}$ is a vector space of functions $f : \mathcal{X} \to \mathbb{R}$ with $\mathcal{F}' \subseteq \mathcal{F}$ and we define

$$\mathcal{H} = \{\mathbf{1}_{\{u:f(u)\leqslant 0\}} : f \in \mathcal{F}'\},$$

then $\mathrm{VC}(\mathcal{H}) \leqslant \dim(\mathcal{F})$.

Let $\mathcal{A} = \left\{\{x : \|x - c\|_2^2 \leqslant r^2\} : c \in \mathbb{R}^d,\, r \in [0, \infty)\right\}$ be the set of all spheres in $\mathbb{R}^d$. Suppose $\mathcal{H} = \{\mathbf{1}_A : A \in \mathcal{A}\}$. Show that

$$\mathrm{VC}(\mathcal{H}) \leqslant d + 2.$$

$\Big[$*Hint: Consider the class of functions* $\mathcal{F}' = \left\{f_{c,r} : c \in \mathbb{R}^d,\, r \in [0, \infty)\right\}$*, where*

$$f_{c,r}(x) = \|x\|_2^2 - 2c^T x + \|c\|_2^2 - r^2\,. \quad \Big]$$

Paper 2, Section II

31J Mathematics of Machine Learning

(a) What is meant by the *subdifferential* $\partial f(x)$ of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at $x \in \mathbb{R}^d$? Write down the subdifferential $\partial f(x)$ of the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \gamma|x|$, where $\gamma > 0$.

Show that x minimises f if and only if $0 \in \partial f(x)$.

What does it mean for a function $f : \mathbb{R}^d \to \mathbb{R}$ to be *strictly convex*? Show that any minimiser of a strictly convex function must be unique.

(b) Suppose we have input–output pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \{-1, 1\}^p \times \{-1, 1\}$ with $p \geqslant 2$. Consider the objective function

$$f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i x_i^T \beta) + \gamma \|\beta\|_1,$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\gamma > 0$. Assume that $(y_i)_{i=1}^n \neq (x_{i1})_{i=1}^n$. Fix $\beta_2, \ldots, \beta_p$ and define

$$\kappa_1 = \sum_{\substack{1 \leqslant i \leqslant n : \\ x_{i1} \neq y_i}} \exp(-y_i \eta_i) \quad \text{and} \quad \kappa_2 = \sum_{i=1}^{n} \exp(-y_i \eta_i),$$

where $\eta_i = \sum_{j=2}^{p} x_{ij} \beta_j$ for $i = 1, \ldots, n$. Show that if $|2\kappa_1 - \kappa_2| \leqslant \gamma$, then

$$\operatorname{argmin}_{\beta_1 \in \mathbb{R}} f(\beta_1, \beta_2, \ldots, \beta_p) = 0.$$

[*You may use any results from the course without proof, other than those whose proof is asked for directly.*]

Paper 4, Section II

30J Mathematics of Machine Learning

Let $D = (x_i, y_i)_{i=1}^n$ be a dataset of n input–output pairs lying in $\mathbb{R}^p \times [-M, M]$ for $M \in \mathbb{R}$. Describe the *random-forest algorithm* as applied to D using decision trees $(\hat{T}^{(b)})_{b=1}^B$ to produce a fitted regression function f_{rf}. [You need not explain in detail the construction of decision trees, but should describe any modifications specific to the random-forest algorithm.]

Briefly explain why for each $x \in \mathbb{R}^p$ and $b = 1, \ldots, B$, we have $\hat{T}^{(b)}(x) \in [-M, M]$.

State the bounded-differences inequality.

Treating D as deterministic, show that with probability at least $1 - \delta$,

$$\sup_{x \in \mathbb{R}^p} |f_{\mathrm{rf}}(x) - \mu(x)| \leqslant M\sqrt{\frac{2\log(1/\delta)}{B}} + \mathbb{E}\Big(\sup_{x \in \mathbb{R}^p} |f_{\mathrm{rf}}(x) - \mu(x)| \Big),$$

where $\mu(x) := \mathbb{E} f_{\mathrm{rf}}(x)$.

$\Big[$*Hint: Treat each* $\hat{T}^{(b)}$ *as a random variable taking values in an appropriate space* $\mathcal{Z}$ *(of functions), and consider a function* G *satisfying*

$$G(\hat{T}^{(1)}, \ldots, \hat{T}^{(B)}) = \sup_{x \in \mathbb{R}^p} |f_{\mathrm{rf}}(x) - \mu(x)|. \Big]$$

Paper 2, Section II

30J Mathematics of Machine Learning

(a) Let $\mathcal{F}$ be a family of functions $f : \mathcal{X} \to \{0, 1\}$. What does it mean for $x_{1:n} \in \mathcal{X}^n$ to be *shattered* by $\mathcal{F}$? Define the *shattering coefficient* $s(\mathcal{F}, n)$ and the *VC dimension* $\mathrm{VC}(\mathcal{F})$ of $\mathcal{F}$.

Let

$$\mathcal{A} = \Big\{ \prod_{j=1}^{d} (-\infty, a_j] : a_1, \ldots, a_d \in \mathbb{R} \Big\}$$

and set $\mathcal{F} = \{\mathbf{1}_A : A \in \mathcal{A}\}$. Compute $\mathrm{VC}(\mathcal{F})$.

(b) State the Sauer–Shelah lemma.

(c) Let $\mathcal{F}_1, \ldots, \mathcal{F}_r$ be families of functions $f : \mathcal{X} \to \{0, 1\}$ with finite VC dimension $v \geqslant 1$. Now suppose $x_{1:n}$ is shattered by $\cup_{k=1}^{r} \mathcal{F}_k$. Show that

$$2^n \leqslant r(n + 1)^v.$$

Conclude that for $v \geqslant 3$,

$$\mathrm{VC}(\cup_{k=1}^{r} \mathcal{F}_k) \leqslant 4v \log_2(2v) + 2 \log_2(r).$$

[You may use without proof the fact that if $x \leqslant \alpha + \beta \log_2(x + 1)$ with $\alpha > 0$ and $\beta \geqslant 3$, then $x \leqslant 4\beta \log_2(2\beta) + 2\alpha$ for $x \geqslant 1$.]

(d) Now let $\mathcal{B}$ be the collection of subsets of $\mathbb{R}^p$ of the form of a product $\prod_{j=1}^{p} A_j$ of intervals A_j, where exactly $d \in \{1, \ldots, p\}$ of the A_j are of the form $(-\infty, a_j]$ for $a_j \in \mathbb{R}$ and the remaining $p - d$ intervals are $\mathbb{R}$. Set $\mathcal{G} = \{\mathbf{1}_B : B \in \mathcal{B}\}$. Show that when $d \geqslant 3$,

$$\mathrm{VC}(\mathcal{G}) \leqslant 2d[2 \log_2(2d) + \log_2(p)].$$

Paper 4, Section II

30J Mathematics of Machine Learning

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 with hypothesis class

$$\mathcal{H} = \{x \mapsto x^T \beta : \beta \in C\},$$

where C is a non-empty closed convex subset of $\mathbb{R}^p$, and logistic loss

$$\ell(h(x), y) = \log_2(1 + e^{-yh(x)}),$$

for $h \in \mathcal{H}$ and $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$.

(i) Show that the objective function f of the optimisation problem is convex.

(ii) Let $\pi_C(x)$ denote the projection of x onto C. Describe the procedure of *stochastic gradient descent* (SGD) for minimisation of f above, giving explicit forms for any gradients used in the algorithm.

(iii) Suppose $\hat{\beta}$ minimises $f(\beta)$ over $\beta \in C$. Suppose $\max_{i=1,\ldots,n} \|x_i\|_2 \leqslant M$ and $\sup_{\beta \in C} \|\beta\|_2 \leqslant R$. Prove that the output $\bar{\beta}$ of k iterations of the SGD algorithm with some fixed step size η (which you should specify), satisfies

$$\mathbb{E} f(\bar{\beta}) - f(\hat{\beta}) \leqslant \frac{2MR}{\log(2)\sqrt{k}}.$$

(iv) Now suppose that the step size at iteration s is $\eta_s > 0$ for each $s = 1, \ldots, k$. Show that, writing β_s for the sth iterate of SGD, we have

$$\mathbb{E} f(\tilde{\beta}) - f(\hat{\beta}) \leqslant \frac{A_2 M^2}{2A_1\{\log(2)\}^2} + \frac{2R^2}{A_1},$$

where

$$\tilde{\beta} = \frac{1}{A_1} \sum_{s=1}^{k} \eta_s \beta_s, \qquad A_1 = \sum_{s=1}^{k} \eta_s \qquad \text{and} \quad A_2 = \sum_{s=1}^{k} \eta_s^2.$$

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