

## Exercise sheet

Lecturer: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

### Part 1. Elements of convex learning problems

**Exercise 1.** (★) Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $f(w) = g(\langle w, x \rangle + y)$  for some  $x \in \mathbb{R}^d, y \in \mathbb{R}$ . Show that: If  $g$  is convex function then  $f$  is convex function.

**Exercise 2.** (★) Let functions  $g_1$  be  $\rho_1$ -Lipschitz and  $g_2$  be  $\rho_2$ -Lipschitz. Then, show that,  $f$  with  $f(x) = g_1(g_2(x))$  is  $\rho_1\rho_2$ -Lipschitz.

**Exercise 3.** (★) Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  with  $f(w) = g(\langle w, x \rangle + y)$   $x \in \mathbb{R}^d$  and  $y \in \mathbb{R}$ . Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a  $\beta$ -smooth function. Then show that  $f$  is a  $(\beta \|x\|^2)$ -smooth.

**Hint::** You may use Cauchy-Schwarz inequality  $\langle y, x \rangle \leq \|y\| \|x\|$

**Exercise 4.** (★) Show that  $f : S \rightarrow \mathbb{R}$  is  $\rho$ -Lipschitz over an open convex set  $S$  if and only if for all  $w \in S$  and  $v \in \partial f(w)$  it is  $\|v\| \leq \rho$ .

**Hint::** You may use Cauchy-Schwarz inequality  $\langle y, x \rangle \leq \|y\| \|x\|$

**Exercise 5.** (★) Let  $g_1(w), \dots, g_r(w)$  be  $r$  convex functions, and let  $f(\cdot) = \max_{j \in [r]} (g_j(\cdot))$ . Show that for some  $w$  it is  $\nabla g_k(w) \in \partial f(w)$  where  $k = \arg \max_j (g_j(w))$  is the index of function  $g_j(\cdot)$  presenting the greatest value at  $w$ .

**Exercise 6.** (★) Consider the regression learning problem  $(\mathcal{H}, \mathcal{Z}, \ell)$  with predictor rule  $h(x) = \langle w, x \rangle$  labeled by some unknown parameter  $w \in \mathcal{W}$ , loss function  $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$ , feature  $x \in \mathcal{X}$ , and target  $y \in \mathbb{R}$ . Let  $\mathcal{W} = \mathcal{X} = \{\omega \in \mathbb{R}^d : |\omega| \leq \rho\}$  for some  $\rho > 0$ .

- (1) Show that the resulting learning problem is Convex-Lipschitz-Bounded learning problem.
- (2) Specify the parameters of Lipschitzness.

**Exercise 7.** (★) If  $f$  is  $\lambda$ -strongly convex and  $u$  is a minimizer of  $f$  then for any  $w$

$$f(w) - f(u) \geq \frac{\lambda}{2} \|w - u\|^2$$

**Hint::** Use the definition, and set  $\alpha \rightarrow 0$ .

The following is given as a homework (Formative assessment 1)

**Exercise 8.** (★) Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a convex and  $\beta$ -smooth function.

(1) Show that for  $v, w \in \mathbb{R}^d$

$$f(v) - f(w) \in \left( \langle \nabla f(w), v - w \rangle, \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2 \right)$$

(2) Show that for  $v, w \in \mathbb{R}^d$  such that  $v = w - \frac{1}{\beta} \nabla f(w)$ , it is

$$\frac{1}{2\beta} \|\nabla f(w)\|^2 \leq f(w) - f(v)$$

(3) Additionally assume that  $f(x) > 0$  for all  $x \in \mathbb{R}^d$ . Show that for  $w \in \mathbb{R}^d$ ,

$$\|\nabla f(w)\| \leq \sqrt{2\beta f(w)}$$

The following is given as a homework (Formative assessment 1)

**Exercise 9.** (★) Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a  $\lambda$ -strongly convex function. Assume that  $w^*$  is a minimizer of  $f$  i.e.

$$w^* = \arg \min_w \{f(w)\}$$

Show that for any  $w \in \mathbb{R}^d$  it holds

$$f(w) - f(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

**Hint:** Use the definition of  $\lambda$ -strongly convex function, properly rearrange it, and ...

**Exercise 10.** (★) Show that the function  $J(x; \lambda) = \lambda \|x\|^2$  is  $2\lambda$ -strongly convex

## Part 2. Stochastic learning

**Exercise 11.** (★) Assume a Bayesian model

$$\begin{cases} z_i | w & \stackrel{\text{ind}}{\sim} f(z_i | w), \quad i = 1, \dots, n \\ w & \sim f(w) \end{cases}$$

and consider that our objective is the discovery of MAP estimate  $w^*$  i.e.

$$w^* = \arg \min_{w \in \Theta} (-\log(L_n(w)) - f(w)) = \arg \min_{w \in \Theta} \left( -\sum_{i=1}^n \log(f(z_i|w)) - \log(f(w)) \right)$$

by using SGD with update

$$w^{(t+1)} = w^{(t)} + \eta_t \left( \frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log(f(z_j|w^{(t)})) + \nabla_w \log(f(w^{(t)})) \right)$$

for some randomly selected set  $\mathcal{J}^{(t)} \subseteq \{1, \dots, n\}^m$  of  $m$  integers from 1 to  $n$  via simple random sampling (SRS) with replacement. Show that

$$\mathbb{E}_{\mathcal{J}^{(t)} \sim \text{simple-random-sampling}} \left( \frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log(f(z_j|w^{(t)})) \right) = \sum_{i=1}^n \nabla_w \log(f(z_i|w^{(t)}))$$


---

**Exercise 12.** (★) Let  $\{v_t; t = 1, \dots, T\}$  be a sequence of vectors. Consider an algorithm producing  $\{w^{(t)}; t = 1, 2, 3, \dots\}$  with

$$\begin{aligned} w^{(1)} &= 0 \\ w^{(t+1)} &= w^{(t)} - \eta v_t \end{aligned}$$

Show that

(1) it is

$$\langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \left( -\|w^{(t+1)} - w^*\|^2 + \|w^{(t)} - w^*\|^2 \right) + \frac{\eta}{2} \|v_t\|^2$$

(2) it is

$$\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \sum_{t=1}^T \left( -\|w^{(t+1)} - w^*\|^2 + \|w^{(t)} - w^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

(3) (continue) it is

$$\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$


---

**Exercise 13.** (★) Let  $\{v_t; t = 1, \dots, T\}$  be a sequence of vectors. Consider an algorithm producing  $\{w^{(t)}; t = 1, 2, 3, \dots\}$  with

$$\begin{aligned} w^{(1)} &= 0 \\ w^{(t+\frac{1}{2})} &= w^{(t)} - \eta v_t \\ w^{(t+1)} &= \arg \min_{w \in \mathcal{H}} \left( \|w - w^{(t+\frac{1}{2})}\| \right) \end{aligned}$$

for  $t = 1, \dots, T$ .

**Hint:** You can use the following Lemma

**(Projection Lemma):** Let  $\mathcal{H}$  be a closed convex set and let  $v$  be the projection of  $w$  onto  $\mathcal{H}$ , i.e.

$$v = \arg \min_{x \in \mathcal{H}} \|x - w\|^2$$

then for every  $u \in \mathcal{H}$  it is

$$\|v - u\|^2 \leq \|w - u\|^2$$

Show that

(1) it is

$$\langle w^{(t)} - w^*, v_t \rangle \leq \frac{1}{2\eta} \left( -\|w^{(t+1)} - w^*\|^2 + \|w^{(t)} - w^*\|^2 \right) + \frac{\eta}{2} \|v_t\|^2$$

(2) it is

$$\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{1}{2\eta} \sum_{t=1}^T \left( -\|w^{(t+1)} - w^*\|^2 + \|w^{(t)} - w^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

(3) (continue) it is

$$\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

**Comment:** Above we show that Lemma ?? from “Handout ??: Gradient descent” holds even when a projection step is included. Hence, even if a projection step is included after the update step of the recursion of GD algorithm or the SGD algorithm the analysis in Section ?? in “Handout ??: Gradient descent” holds. Hence, even if a projection step is included after the update step of the recursion of SGD algorithm or the SGD algorithm the analysis in Section ?? in “Handout ??: Stochastic gradient descent” holds.

The following is given as a homework (Formative assessment 2)

**Exercise 14.** (\*) <sup>1</sup>Consider the binary classification problem with inputs  $x \in \mathcal{X}$  where  $\mathcal{X} := \{x \in \mathbb{R}^d : \|x\|_2 \leq L\}$  for some given value  $L > 0$ , target  $y \in \mathcal{Y}$  where  $\mathcal{Y} := \{-1, +1\}$ , and prediction rule  $h_w : \mathbb{R}^d \rightarrow \{-1, +1\}$  with

$$\begin{aligned} (1) \quad h_w(x) &= \text{sign}(w^\top x) \\ (2) \quad &= \text{sign}\left(\sum_{j=1}^d w_j x_j\right) \end{aligned}$$

<sup>1</sup>We use standard notation

$$\text{sign}(\xi) = \begin{cases} -1, & \text{if } \xi < 0 \\ +1, & \text{if } \xi > 0 \end{cases}$$

$\pm 1$  means either  $-1$  or  $+1$ ,  $\mathbb{R}_+ := (0, +\infty)$ , and  $\|x\|_2 := \sqrt{\sum_{j=1}^d (x_j)^2}$  for the Euclidean distance.

Let the hypothesis class is

$$(3) \quad \mathcal{H} = \left\{ x \rightarrow w^\top x : \forall w \in \mathbb{R}^d \right\}$$

In other words, the hypothesis  $h_w \in \mathcal{H}$  is parametrized by  $w \in \mathbb{R}^d$ , it receives an input vector  $x \in \mathcal{X} := \mathbb{R}^d$  and it returns the label  $y = \text{sign}(w^\top x) \in \mathcal{Y} := \{\pm 1\}$  where

$$\text{sign}(\xi) = \begin{cases} -1, & \text{if } \xi < 0 \\ +1, & \text{if } \xi > 0 \end{cases}$$

Consider a loss function  $\ell : \mathbb{R}^d \rightarrow \mathbb{R}_+$  with

$$(4) \quad \ell(w, z = (x, y)) = \max(0, 1 - yw^\top x) + \lambda \|w\|_2^2$$

for some given value  $\lambda > 0$ .

Assume there is available a dataset of examples  $S_n = \{z_i = (x_i, y_i) ; i = 1, \dots, n\}$  of size  $n$ .

Do the following:

- (1) Show that the function  $f : \mathbb{R} \rightarrow \mathbb{R}_+$  with  $f(x) = \max(0, 1 - x)$  is convex in  $\mathbb{R}$ ; and show that the loss (4) is convex.

**Hint::** You may use Proposition ?? from Handout ?? : Elements of convex learning problems.

- (2) Show that the loss  $\ell(w, z)$  for  $\lambda = 0$  (4) is  $L$ -Lipschitz (with respect to  $w$ ) when  $x \in \mathcal{X}$  where  $\mathcal{X} := \{x \in \mathbb{R}^d : \|x\|_2 \leq L\}$ .

**Hint::** You may use the definition of Lipschitz function. Without loss of generality, you can consider any  $w_1 \in \mathbb{R}^d$  and  $w_2 \in \mathbb{R}^d$  such that  $1 - yw_2^\top x \leq 1 - yw_1^\top x$ , and then take cases  $1 - yw_2^\top x > \text{or} < 0$  and  $1 - yw_1^\top x > \text{or} < 0$  to deal with the max.

- (3) Construct the set of sub-gradients  $\partial f(x)$  for  $x \in \mathbb{R}$  of the function  $f : \mathbb{R} \rightarrow \mathbb{R}_+$  with  $f(x) = \max(0, 1 - x)$ . Show that the vector  $v$  with

$$v = \begin{cases} 2\lambda w, & yw^\top x > 1 \\ 2\lambda w, & yw^\top x = 1 \\ -yx + 2\lambda w, & yw^\top x < 1 \end{cases}$$

is  $v \in \partial_w \ell(w, z = (x, y))$ , aka a sub-gradient of  $\ell(w, z = (x, y))$  at  $w$ , for any  $w \in \mathbb{R}^d$ .

- (4) Write down the algorithm of online AdaGrad (Adaptive Stochastic Gradient Descent) with learning rate  $\eta_t > 0$ , batch size  $m$ , and termination criterion  $t > T_{\max}$  for some  $T_{\max} > 0$  in order to discover  $w^*$  such as

$$(5) \quad w^* = \arg \min_{w: h_w \in \mathcal{H}} (\mathbb{E}_{z \sim g} (\ell(w, z = (x, y))))$$

The formulas in your algorithm should be implemented for the above learning problem and tailored to 1, 3, and 4.

- (5) Use the R code given below in order to generate the dataset of observed examples  $S_n = \{z_i = (x_i, y_i)\}_{i=1}^n$  that contains  $n = 10^6$  examples with inputs  $x$  of dimension  $d = 2$ . Consider  $\lambda = 0$ . Use a seed  $w^{(0)} = (0, 0)^\top$ .

- (a) By using appropriate values for  $m$ ,  $\eta_t$  and  $T_{\max}$ , code in R the algorithm you designed in part 4, and run it.
- (b) Plot the trace plots for each of the dimensions of the generated chain  $\{w^{(t)}\}$  against the iteration  $t$ .
- (c) Report the value of the output  $w_{\text{adaGrad}}^*$  (any type) of the algorithm as the solution to (5).
- (d) To which cluster  $y$  (i.e.,  $-1$  or  $1$ )  $x_{\text{new}} = (1, 0)^\top$  belongs?

```
# R code. Run it before you run anything else
#
data_generating_model <- function(n,w) {
  z <- rep( NaN, times=n*3 )
  z <- matrix(z, nrow = n, ncol = 3)
  z[,1] <- rep(1,times=n)
  z[,2] <- runif(n, min = -10, max = 10)
  p <- w[1]*z[,1] + w[2]*z[,2] p <- exp(p) / (1+exp(p))
  z[,3] <- rbinom(n, size = 1, prob = p)
  ind <- (z[,3]==0)
  z[ind,3] <- -1
  x <- z[,1:2]
  y <- z[,3]
  return(list(z=z, x=x, y=y))
}
n_obs <- 1000000
w_true <- c(-3,4)
set.seed(2023)
out <- data_generating_model(n = n_obs, w = w_true)
set.seed(0)
z_obs <- out$z #z=(x,y)
x <- out$x
y <- out$y
#z_obs2=z_obs
#z_obs2[z_obs[,3]==-1,3]=0
#w_true <- as.numeric(glm(z_obs2[,3]~ 1+ z_obs2[,2],family = "binomial"
)$coefficients)
```