

## Exercise sheet

Lecturer: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

### Part 1. Stochastic learning

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

**Solution.** Let  $u, v \in \mathbb{R}^d$  and  $\alpha \in [0, 1]$ . It is

$$\begin{aligned}
 f(\alpha u + (1 - \alpha)v) &= g(\langle \alpha u + (1 - \alpha)v, x \rangle + y) \\
 &= g(\alpha \langle u, x \rangle + (1 - \alpha) \langle v, x \rangle + y) \\
 &= g(\alpha (\langle u, x \rangle + y) + (1 - \alpha) (\langle v, x \rangle + y)) & y &= \alpha y + (1 - \alpha)y \\
 &\leq \alpha g(\langle u, x \rangle + y) + (1 - \alpha) g(\langle v, x \rangle + y) & (g \text{ is convex}) \\
 &= \alpha f(u) + (1 - \alpha) f(v)
 \end{aligned}$$

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

**Solution.**

$$\begin{aligned}
 |f(w_1) - f(w_2)| &= |g_1(g_2(w_1)) - g_1(g_2(w_2))| \\
 &\leq \rho_1 |g_2(w_1) - g_2(w_2)| \\
 &\leq \rho_1 \rho_2 |w_1 - w_2|
 \end{aligned}$$

**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  $f$  is a  $(\beta \|x\|^2)$ -smooth.

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

$$f(v) = g(\langle w, x \rangle + y)$$

$$\leq g(\langle w, x \rangle + y) + g'(\langle w, x \rangle + y) \langle v - w, x \rangle + \frac{\beta}{2} (\langle v - w, x \rangle)^2 \quad (g \text{ is smooth})$$

$$\leq g(\langle w, x \rangle + y) + g'(\langle w, x \rangle + y) \langle v - w, x \rangle + \frac{\beta}{2} (\|v - w\| \|x\|)^2 \quad (\text{Cauchy-Schwarz inequality})$$

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

**Exercise 4.**  $(\star) f : S \rightarrow \mathbb{R}$  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  $\langle y, x \rangle \leq \|y\| \|x\|$

**Solution.**  $\implies$  Let  $f : S \rightarrow \mathbb{R}$  be  $\rho$ -Lipschitz over convex set  $S$ ,  $w \in S$  and  $v \in \partial f(w)$ .

- Since  $S$  is open we get that there exist  $\epsilon > 0$  such as  $u := w + \epsilon \frac{v}{\|v\|}$  where  $u \in S$ . So  $\langle u - w, v \rangle = \epsilon \|v\|$  and  $\|u - w\| = \epsilon$ .
- From the subgradient definition we get

$$f(u) - f(w) \geq \langle u - w, v \rangle = \epsilon \|v\|$$

- From the Lipschitzness of  $f(\cdot)$  we get

$$f(u) - f(w) \leq \rho \|u - w\| = \rho \epsilon$$

Therefore  $\|v\| \leq \rho$ .

*Proof.*  $\Leftarrow$  It is for all  $w \in S$  and  $v \in \partial f(w)$  it is  $\|v\| \leq \rho$ . □

- For any  $u \in S$ , it is

$$\begin{aligned} f(w) - f(u) &\leq \langle v, w - u \rangle && (\text{because } v \in \partial f(w)) \\ (1) \quad &\leq \|v\| \|w - u\| && \text{by Cauchy-Schwarz inequality} \\ &\leq \rho \|w - u\| && \text{because } \|v\| \leq \rho \end{aligned}$$

- Similarly it results  $u, w \in S$

$$f(w) - f(u) \leq \langle v, u - w \rangle \|v\| \leq \|v\| \|u - w\| \leq \rho \|u - w\|$$

from (1) because  $w, u$  can be swaped in (1) as they both are any values in  $S$ .

**Exercise 5.**  $(\star)$  Let  $g_1(w), \dots, g_r(w)$  be  $r$  convex functions, and let  $f(\cdot) = \max_{\forall j} (g_j(\cdot))$ . 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  $w$ .

Since  $g_j$  is convex, for all  $u$

$$g_j(u) \geq g_j(w) + \langle u - w, \nabla g_j(w) \rangle$$

However  $f(u) = \max_{\forall j} (g_j(u)) \geq g_j(u)$  for any  $j$ , and  $f(w) = g_k(w)$  at  $w$ . Then

$$f(u) \geq f(w) + \langle u - w, \nabla g_k(w) \rangle$$

Then by the definition of the sub-gradient  $\nabla g_k(w) \in \partial f(w)$

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

**Exercise 6.** (★) 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

$$(2) \quad h_w(x) = \text{sign}(w^\top x)$$

$$(3) \quad = \text{sign}\left(\sum_{j=1}^d w_j x_j\right)$$

Let the hypothesis class of prediction rules be

$$\mathcal{H} = \{x \rightarrow w^\top x : \forall w \in \mathbb{R}^d\}$$

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  $y = \text{sign}(w^\top x) \in \mathcal{Y} := \{\pm 1\}$ .

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 tasks.

**Hint-1::** We denote

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

**Hint-2::** The notation  $\pm 1$  means either  $-1$  or  $+1$ .

**Hint-3::** We define  $\mathbb{R}_+ := (0, +\infty)$

**Hint-4::** We denote  $\|x\|_2 := \sqrt{\sum_{\forall j} (x_j)^2}$  the Euclidean distance.

- (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 Example 13 from Handout 1.

- (2) Show that the loss  $\ell$  (4) is  $L$ -Lipschitz 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 have to be tailored to 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 suitable values for  $m$ ,  $\eta_t$  and  $T_{\max}$ , code 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)

```