Homework 1: Stochastic learning: Stochastic Gradient Descent

Lecturer: Georgios P. Karagiannis   georgios.karagiannis@durham.ac.uk

As formative assessment, submit the solutions to Exercise 1.2, 1.3, and 1.4.

Exercise 1. ($\star\star\star$) 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 $h_w : \mathbb{R}^d \to \{-1, +1\}$ with

$$(0.1) \qquad h_w(x) = \text{sign}\left(w^\top x\right)$$

$$(0.2) \qquad = \text{sign}\left(\sum_{j=1}^{d} w_j x_j\right)$$

Let the hypothesis class of prediction rules be

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

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

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

$$(0.3) \qquad \ell(w, z = (x, y)) = \max\left(0, 1 - y w^\top x\right) + \lambda \|w\|_2^2$$

for some given value $\lambda > 0$.

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

Do the following tasks.

Hint-1:: We denote

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

Hint-2:: The notation ± 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 $f : \mathbb{R} \to \mathbb{R}_+$ with $f(x) = \max(0, 1 - x)$ is convex in $\mathbb{R}$; and show that the loss (0.3) is convex.

    Hint:: You may use Example 13 from Handout 1.

(2) Show that the loss $\ell(w, z)$ for $\lambda = 0$ (0.3) is L-Lipschitz (with respect to w) when $x \in \mathcal{X}$ where $\mathcal{X} := \{x \in \mathbb{R}^d : \|x\|_2 \leq L\}$.

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 > $ or < 0 and $1 - yw_1^\top x > $ 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 $f : \mathbb{R} \to \mathbb{R}_+$ 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

(0.4)
$$
w^* = \arg\min_{\forall w : h_w \in \mathcal{H}} \left(\mathbb{E}_{z \sim g} \left(\ell(w, z = (x, y)) \right) \right)
$$

The formulas in your algorithm have to be tailored to 0.3.

(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, η_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^*_{adaGrad} (any type) of the algorithm as the solution to (0.4).

   (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)
```

Solution.

(1) $f_1(x) = 0$ is convex, $f_2(x) = 1 - x$ is convex, hence from the example in Handout 1, $f(x) = \max(f_1(x), f_2(x))$ is convex as well. Regarding the loss function, we just have $f_2(w) = 1 - yx^\top w$ which is convex as a composition due to linearity.

(2) Given a fixed example $(x,y) \in \{x \in \mathbb{R}^d : \|x'\|_2 \leq R\} \times \{-1,1\}$.

Assume $w_1, w_2 \in \mathbb{R}^d$. Let $\ell_i = \max\{0, 1 - yx^\top w_i\}$, for $i = 1,2$. It suffices to show that $|\ell_1 - \ell_2|_2 \leq R|w_1 - w_2|_2$. I take cases

Case-1: Assume $yx^\top w_1 \geq 1$ and $yx^\top w_2 \geq 1$ then $|\ell_1 - \ell_2|_2 = 0 \leq R|w_1 - w_2|_2$

Case-2: Assume that at least one of $yx^\top w_1 < 1$ or $yx^\top w_2 < 1$ but not both is true. Assume without loss of generality that $1 - yx^\top w_1 < 1 - yx^\top w_2$. Then

$$
\begin{aligned}
|\ell_1 - \ell_2|_2 &= \ell_1 - \ell_2 \\
&= 1 - yx^\top w_1 - \max\left(0, 1 - yx^\top w_2\right) \\
&\leq 1 - yx^\top w_1 - \left(1 - yx^\top w_2\right) \\
&= yx^\top (w_2 - w_1) \\
&\leq y \left\| x^\top \right\|_2 \|w_1 - w_2\|_2 \quad \text{because} \quad a^\top b \leq \|a\| \|b\|
\end{aligned}
$$

(3) It is

$$
f(x) = \max(0, 1 - x) = \begin{cases} 0 & x > 1 \\ 0 & x = 1 \\ 1 - x & x < 1 \end{cases}
$$

- For $x > 1$, f is differentiable so $\partial f(x) = \{f'(x)\} = \{0\}$.
- For $x < 1$, f is differentiable so $\partial f(x) = \{f'(x)\} = \{-1\}$.
- For $x = 1$, f is not differentiable. By definition I have that v is subgradient of $f(x)$ at $x = 0 \in S$ if

$$\forall u \in \mathbb{R}, \quad f(u) \geq f(x) + \langle u - x, v \rangle$$

So, for $u \geq 1$, it is $0 \geq (u - 1)v \implies v \leq 0$, and for $u < 1$ it is $(1 - u) \geq (u - 1)v \implies v \geq -1$. Hence the common space is $v \in [0, 1]$ So $\partial f(x) = [0, 1]$. Hence,

$$
\partial f(x) = \begin{cases} 0, & x > 1 \\ [0, 1], & x = 1 \\ -1, & x < 1 \end{cases}
$$

Now regarding the loss $\partial_w \ell(w, z = (x, y))$
- for $yw^\top x > 1$ it is differentiable so $\nabla_w \ell(w, z = (x, y)) = \nabla_w \left(0 + \lambda \sum_{j=1}^d w_j^2\right) = 2\lambda w$; as

$$\frac{\mathrm{d}}{\mathrm{d}w_j} \sum_{j'=1}^d w_{j'}^2 = 2\lambda w_j$$

- for $yw^\top x > 1$ it is differentiable so $\nabla_w \ell(w, z = (x, y)) = \nabla_w \left(1 - yw^\top x + \lambda \sum_{j=1}^d w_j^2\right) = yx + 2\lambda w$ as

$$\frac{\mathrm{d}}{\mathrm{d}w_j} \left(1 - yw^\top x\right) = \frac{\mathrm{d}}{\mathrm{d}w_j} \left(1 - y \sum_{j'=1}^d w_{j'} x_{j'}\right) = -yx_j$$

- for $yw^\top x = 1$, $v = 0$ satisfies the definition of the sub-gradient

$$\forall u, \quad f(u) \geq \cancel{f(w)}^{\,0} + \langle u - w, v \rangle$$

$$\max\left(0, 1 - yu^\top x\right) \geq 0 + (u - w)^\top 0$$

So

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

$$= \partial\left(\max\left(0, 1 - yw^\top x\right)\right) + \partial\left(\lambda \|w\|_2^2\right)$$

$$= \partial\left(\max\left(0, 1 - yw^\top x\right)\right) + \nabla\left(\lambda \|w\|_2^2\right)$$

$$0 + 2\lambda w$$

but $\partial\left(\lambda \|w\|_2^2\right) = \left\{\nabla\left(\lambda \|w\|_2^2\right)\right\}$ because $\lambda \|w\|_2^2$ is differentiable. Hence

$$\partial \ell\left(w, z = (x, y)\right) = 0 + 2\lambda w$$

Hence

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

(4)

Algorithm. *For $t = 1, 2, 3, \ldots$ iterate:*

(a) *get a random sub-sample $\left\{\tilde{z}_j^{(t)}; j = 1, \ldots, m\right\}$ of size m with or without replacement from the complete data-set $\mathcal{S}_n$.*

(b) *compute*

$$w^{(t+1)} = w^{(t)} - \eta_t v_t,$$

where $v_t = \frac{1}{m} \sum_{j=1}^m \tilde{v}_{t,j}$ and

$$\tilde{v}_{t,j} = \begin{cases} 2\lambda w^{(t)}, & \tilde{y}_j^{(t)} \left(w^{(t)}\right)^\top \tilde{x}_j^{(t)} > 1 \\ 2\lambda w^{(t)}, & \tilde{y}_j^{(t)} \left(w^{(t)}\right)^\top \tilde{x}_j^{(t)} = 1 \\ -\frac{1}{m} \tilde{y}_j^{(t)} \tilde{x}_j^{(t)} + 2\lambda w^{(t)}, & \tilde{y}_j^{(t)} \left(w^{(t)}\right)^\top \tilde{x}_j^{(t)} < 1 \end{cases}$$

where I denote $\tilde{z}_j^{(t)} = \left(\tilde{x}_j^{(t)}, \tilde{y}_j^{(t)}\right)$.

(c) *terminate if a termination criterion is satisfied*

(5)

(a) The R code can be found in the link `https://raw.githubusercontent.com/georgios-stats/` `Machine_Learning_and_Neural_Networks_III_Epiphany_2023/main/Exercises/supplementary/` `q6.R`

(b) The figures are presented below
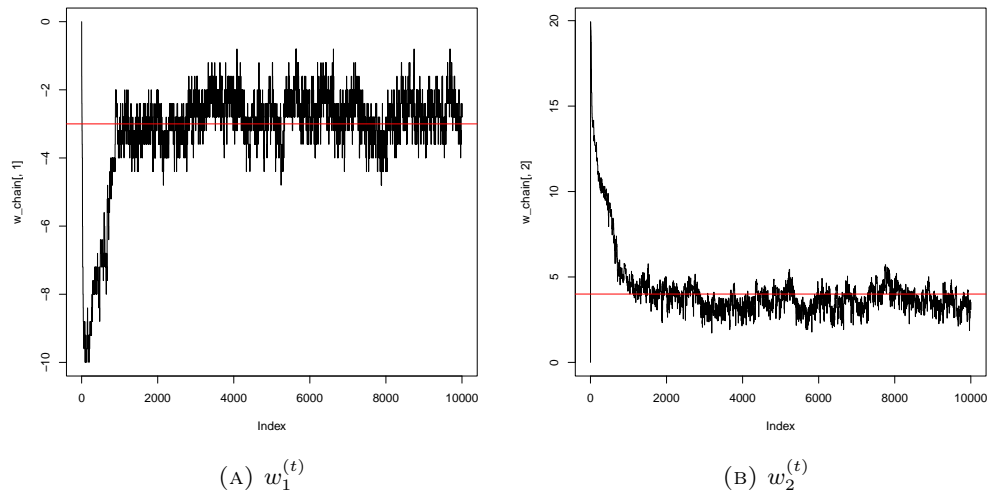
(A) $w_1^{(t)}$ (B) $w_2^{(t)}$

FIGURE 0.1. trace plots

(c) I found $w = (-2.674615, 3.205785)$

(d)   It belongs to -1