

# Modern intro to Online Learning – Exercise Solutions

Ludovic Schwartz

# Contents

<b>1</b>	<b>What is online learning ?</b>	<b>2</b>
	Exercise 1. . . . .	2
<b>2</b>	<b>Online Subgradient Descent</b>	<b>4</b>
	Exercise 1. . . . .	4
	Exercise 2. . . . .	4
	Exercise 3. . . . .	5
	Exercise 4. . . . .	6

# Chapter 1

## What is online learning ?

### Exercise 1.1

Extend the previous algorithm and analysis to the case when the adversary selects a vector  $y_t \in \mathbb{R}^d$  such that  $\|y_t\|_2 \leq 1$ , the algorithm guesses a vector  $x_t \in \mathbb{R}^d$ , and the loss function is  $\|x_t - y_t\|_2^2$ . Show an upper bound to the regret logarithmic in  $T$  and that does not depend on  $d$ . Among the other things, you will probably need the Cauchy-Schwarz inequality :  $\langle x, y \rangle \leq \|x\|_2 \|y\|_2$ .

**Solution :** The natural extension of the previous algorithm is to pick the  $x_t$  which minimizes the cumulated loss up to time  $t - 1$ . We also define  $x_t^*$  the optimal comparator at times  $t$  :

$$x_t = x_{t-1}^* = \arg \min_{x \in \mathbb{R}} \sum_{s=1}^{t-1} \|x - y_s\|^2$$

We can once again explicitly compute the value of  $x_t$ . Indeed, if we define  $F_t(x) := \sum_{s=1}^t \|x - y_s\|^2$ ,  $F_t$  is a strictly convex function and reaches its minimum where its gradient vanishes. A simple computation then gives  $\nabla F_t(x) = 0 \Leftrightarrow \sum_{s=1}^t 2(x - y_s) = 0 \Leftrightarrow x = \frac{1}{t} \sum_{s=1}^t y_s$ . In particular, we have again  $x_t = \frac{1}{t-1} \sum_{s=1}^{t-1} y_s$  and we can notice that  $\|x_t\| \leq 1$  at all time.

Now, by lemma 1.2 with the loss  $\ell_t(x) = \|x - y_t\|^2$ , we have :

$$\forall T, \sum_{t=1}^T \|x_T^* - y_t\|^2 \geq \sum_{s=1}^T \|x_s^* - y_t\|^2$$

We can now complete the proof :

$$\begin{aligned}
R_T &= \sum_{t=1}^T \|x_t - y_t\|^2 - \min_{x \in \mathbb{R}} \sum_{t=1}^T \|x - y_t\|^2 \\
&= \sum_{t=1}^T \|x_{t-1}^* - y_t\|^2 - \sum_{t=1}^T \|x_T - y_t\|^2 \\
&\leq \sum_{t=1}^T \|x_{t-1}^* - y_t\|^2 - \sum_{t=1}^T \|x_t^* - y_t\|^2 \\
&= \sum_{t=1}^T \langle x_{t-1}^* + x_t^* - 2y_t, x_{t-1}^* - x_t^* \rangle \\
&\stackrel{\text{(C.S)}}{\leq} \sum_{t=1}^T \|x_{t-1}^* + x_t^* - 2y_t\| \cdot \|x_{t-1}^* - x_t^*\| \\
&\leq \sum_{t=1}^T 4 \|x_{t-1}^* - x_t^*\|
\end{aligned}$$

Where the first inequality uses lemma 1.2, the second one uses Cauchy-Schwarz and the third uses that  $\forall t, \|x_t^*\| \leq 1$  and  $\|y_t\| \leq 1$ . Now we notice that

$$\begin{aligned}
\|x_{t-1}^* - x_t^*\| &= \left\| \frac{1}{t-1} \sum_{s=1}^{t-1} y_s - \frac{1}{t} \sum_{s=1}^t y_s \right\| \\
&= \left\| \frac{1}{t(t-1)} \sum_{s=1}^{t-1} y_s + \frac{1}{t} y_t \right\| \\
&\leq \frac{1}{t(t-1)} \sum_{s=1}^{t-1} \|y_s\| + \frac{1}{t} \|y_t\| \\
&\leq \frac{2}{t}
\end{aligned}$$

Now we plug everything together :

$$R_T \leq 4 \cdot \sum_{t=1}^T \|x_{t-1}^* - x_t^*\| \leq 8 \cdot \sum_{t=1}^T \frac{1}{t} \leq 8 + 8 \log T$$

□

## Chapter 2

# Online Subgradient Descent

### Exercise 2.1

Prove that  $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T} - 1$

**Solution :** We have :

$$\begin{aligned}\sum_{t=1}^T \frac{1}{\sqrt{t}} &= 1 + \sum_{t=2}^T \frac{1}{\sqrt{t}} \\ &= 1 + \sum_{t=2}^T \int_{t-1}^t \frac{1}{\sqrt{t}} du \\ &\leq 1 + \sum_{t=2}^T \int_{t-1}^t \frac{1}{\sqrt{u}} du \\ &\leq 1 + \int_1^T \frac{1}{\sqrt{u}} du \\ &= 1 + 2\sqrt{T} - 2 = 2\sqrt{T} - 1\end{aligned}$$

□

### Exercise 2.2

Using the inequality in the previous exercise, prove that a learning rate  $\propto \frac{1}{\sqrt{t}}$  gives rise to a regret only a constant multiplicative factor worse than the one in (2.1) ( $R_T \leq DL\sqrt{T}$ ).

**Solution :** We start at the result of theorem 2.13 :

$$R_T \leq \frac{D^2}{2\eta_T} + \sum_{t=1}^T \frac{\eta_t}{2} \|g_t\|^2$$

Then we bound for any  $t$ ,  $\|g_t\|^2 \leq L^2$  and set  $\eta_t = \alpha \frac{1}{\sqrt{t}}$  with  $\alpha > 0$  to be determined later. We have

$$\begin{aligned}
R_T &\leq \frac{D^2\sqrt{T}}{2\alpha} + \sum_{t=1}^T \frac{\alpha}{2\sqrt{t}} L^2 \\
&= \frac{D^2\sqrt{T}}{2\alpha} + \frac{\alpha L^2}{2} \sum_{t=1}^T \frac{1}{\sqrt{t}} \\
&\leq \frac{D^2\sqrt{T}}{2\alpha} + \alpha L^2 \sqrt{T} \\
&= \sqrt{T} \left( \frac{D^2}{2\alpha} + \alpha L^2 \right) \\
&= DL\sqrt{2T}
\end{aligned}$$

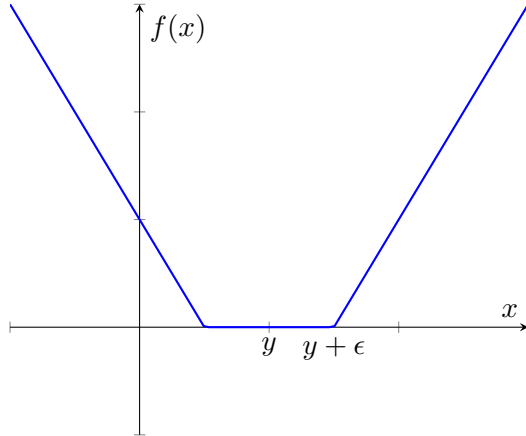
Where the third line uses the result of the previous exercise and the last line uses the choice  $\alpha = \sqrt{\frac{D^2}{2L^2}}$ . We remark that we are only a factor  $\sqrt{2}$  worse than the bound obtained with a fixed  $\eta$ .

□

### Exercise 2.3

Calculate the subdifferential set of the  $\epsilon$ -insensitive loss :  
 $f(x) = \max(|x - y| - \epsilon, 0)$

**Solution :** We start by a drawing of the function :



Then the subdifferential is :

$$\partial f(x) = \begin{cases} \{0\} & \text{if } x \in ]y - \epsilon, y + \epsilon[ \\ \{-1\} & \text{if } x \in ]-\infty, y - \epsilon[ \\ \{1\} & \text{if } x \in ]y + \epsilon, \infty[ \\ [0, 1] & \text{if } x = y + \epsilon \\ [-1, 1] & \text{if } x = y - \epsilon \end{cases}$$

□

#### Exercise 2.4

Using the definition of subgradient, find the subdifferential set of  $f(x) = \|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$ ,  $x \in \mathbb{R}^d$

**Solution :** We start by treating the case  $x \neq 0$ . Then  $f$  is actually differentiable in  $x$  and we have :

$$\frac{\partial f}{\partial x_i}(x) = \frac{x_i}{\sqrt{\sum_{i=1}^d x_i^2}} = \frac{x_i}{\|x\|_2}$$

In that case, we simply have :

$$\partial f(x) = \{\nabla f(x)\} = \left\{ \frac{x}{\|x\|_2} \right\}$$

For the case  $x = 0$  we use the definition of a subgradient,  $g$  is a subgradient of  $f$  at the point  $x = 0$  if and only if

$$\forall y \in \mathbb{R}^d, \|y\|_2 \geq \|0\|_2 + \langle g, y \rangle = \langle g, y \rangle$$

By Cauchy Schwarz Inequality, we have :

$$\forall g, y \in \mathbb{R}^d, \langle g, y \rangle \leq \|g\|_2 \cdot \|y\|_2$$

So any  $g$  of norm smaller than 1 will be a subgradient of  $f$  at  $x = 0$ . Conversely, if  $\|g\|_2 > 1$ , then we have for  $y = g$ ,  $\langle g, y \rangle = \langle y, y \rangle = \|y\|_2^2 > \|y\|_2$  and  $g$  will not be a subgradient of  $f$  at  $x = 0$ . To conclude, we have that :

$$\partial f(x) = \begin{cases} \left\{ \frac{x}{\|x\|_2} \right\} & \text{if } x \neq 0 \\ \{g, \|g\|_2 \leq 1\} & \text{if } x = 0 \end{cases}$$

□