Modern intro to Online Learning – Exercise Solutions

Ludovic Schwartz

Contents

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 $\|y\|_2 \leq 1$, 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 :

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

We can once again explicitely compute the value of x_t. Indeed, if we define $F_t(x) := \sum_{s=1}^{t} \|x - y_t\|^2$, 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 :

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

We can now complete the proof :

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

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

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

Now we plug everything together :

$$
R_T \leq 4 \cdot \sum_{t=1}^{T} \left\|x^*_{t-1} - x^*_t\right\| \leq 8 \cdot \sum_{t=1}^{T} \frac{1}{t} \leq 8 + 8 \log T
$$

◻

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

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

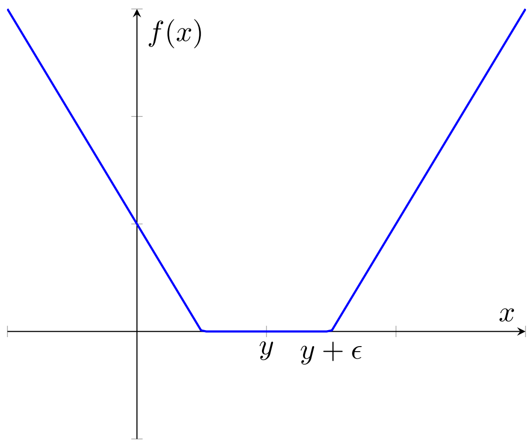
Where the third line uses the result of the previous exercise and the last lines uses the choice $\alpha = \sqrt{\frac{D^2}{2L^2}}$. We remark that we are only a factor $\sqrt{2}$ worse than the bound obtained with a fixed η.

$\square$

Exercise 2.3

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

Solution : We start by a drawing of the function :



Then the subdifferential is :

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

$\boxtimes$

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 :

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

$\boxtimes$

Exercise 2.5

Consider Projected Online Subgradient Descent for the example 2.10 on the failure of Follow-the-Leader: Can we use it on that problem ? Would it guarantee sublinear regret ? How would the behaviour of the algorithm differ from FTL ?

Solution : The setting of Example 2.10 is the following : Let $V = [-1, 1]$ and consider the sequence of losses $\ell_t(x) = z_t x + i_v(x)$, where

$$z_t = \begin{cases} -0.5 & \text{if } t = 1 \\ (-1)^t & \text{if } t > 1 \end{cases}$$

We can apply Projected Online Subgradient Descent in that problem. In particular, since we only do prediction inside of V, we can consider the loss function $\ell_t(x) = z_t x$. We have :

$$\nabla \ell_t(x) = z_t$$

Now we can verify that :

$$\forall x \in V, \|\nabla \ell_t(x)\| \leq 1 := L$$

and that :

$$D := \sup_{x,y \in V} \|x - y\| = 2 < \infty$$

Theorem 2.13 applies and we get a bound of order $DL\sqrt{T} = 2\sqrt{T}$, that is to say, sublinear regret. The main difference with Follow the Leader is that Projected Online Subgradient Decent will pick points in the interior of V and do small updates on them while Follow the leader picks extremal points and moves a lot between each prediction.

$\square$

Chapter 4

Beyond $\sqrt{T}$ Regret

Exercise 4.1

Prove that OSD in Example 4.11 with $x_1 = 0$ is exactly the Follow-the-Leader strategy for that particular problem.

Solution : In example 4.11, we have $\ell_t(x) = (x - y_t)^2$, $\eta_t = \frac{1}{2t}$. We can explicitely compute the predictions of OSD :

$$x_1 = 0$$

and

$$\begin{aligned} x_{t+1} &= x_t - \eta_t \nabla \ell_t(x_t) \\ &= x_t - \frac{1}{t}(x_t - y_t) \\ &= \frac{t-1}{t} x_t + \frac{1}{t} \end{aligned}$$

An elementary induction then gives $x_{t+1} = \frac{1}{t}\sum_{s=1}^{t} y_s$ which is exactly the strategy of Follow-the-Leader on that problem. □

Exercise 4.4

Prove that the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q \geq 1$

Solution : Let $p, q \geq 1$. By absolute homogeneity of the norms, it suffices to prove that $\forall y \in \mathbb{R}^d$, $\|y\|_q = 1 \implies \|y\|_p^* = 1$.

Let $y \in \mathbb{R}^d$ such that $\|y\|_q = 1$, we have

$$\|y\|_p^* = \max_{\|x\|_p = 1} \langle x, y \rangle \leq \|x\|_p \|y\|_q = 1$$

by Hölder's inequality with equality for some value of x. For completeness, we reprove the equality case of Hölder's inequality here. If $1 < p < \infty$, we define $x_i = sgn(y_i) \cdot |y_i|^{\frac{q}{p}}$ (With the convention $sgn(0) = 0$) We have :

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{d} |y_i|^{p\cdot\frac{q}{p}}\right)^{\frac{1}{p}} = \|y\|_q^{\frac{q}{p}} = 1$$

And

$$\langle x, y\rangle = \sum_{i=1}^{d} x_i \cdot y_i = \sum_{i=1}^{d} |y_i|^{1+\frac{q}{p}} = \sum_{i=1}^{d} |y_i|^q = \|y\|_q^q = 1$$

If $p = 1$, let $i^* \in \arg\max_i |y_i|$ and $x_i = sgn(y_{i^*})\mathbb{I}_{i=i^*}$. We have $\|x\|_1 = 1$, and $\langle x, y\rangle = |y_{i^*}| = \|y\|_\infty = 1$

If $p = \infty$, let $x_i = sgn(y_i)$(with the convention $sgn(0) = 0$). We have $\|x\|_\infty = 1$ and $\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i = \sum_{i=1}^{d} |y_i| = \|y\|_1 = 1$

□

Exercise 4.5

Show that using online subgradient descent on a bounded domain V with the learning rates $\eta_t = O(1/t)$ with Lipschitz, smooth, and strongly convex functions, you can get $O(\log(1 + L^*))$ bounds.

Solution : We assume that every loss function ℓ_t is K Lipschitz, μ strongly convex and M smooth. Without loss of generality, we assume that the minimum of each loss function on V is equal to 0.

We set the learning rate as $\eta_t = \frac{1}{\sum_{s=1}^{t} \mu_s} = \frac{1}{\mu t}$. This learning rates verify :

$$\frac{1}{2\eta_1} - \frac{\mu}{2} = 0$$
$$\frac{1}{2\eta_t} - \frac{\mu}{2} = \frac{1}{2\eta_{t-1}}$$

Then, using again Lemma 2.23, and summing from 1 to T, we get :

$$\begin{aligned}\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) &\le \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|x_t - u\|_2^2 - \frac{1}{2\eta_t}\|x_{t+1} - u\|_2^2 - \frac{\mu}{2}\|x_t - u\|_2^2 + \frac{\eta_t}{2}\|g_t\|_2^2\right)\\
&= -\frac{1}{2\eta_1}\|x_2 - u\|_2^2 + \sum_{t=2}^{T}\left(\frac{1}{2\eta_{t-1}}\|x_t - u\|_2^2 - \frac{1}{2\eta_t}\left\|x_{t+1} - u_2^2\right\|\right) + \sum_{t=1}^{T}\frac{\eta_t}{2}\|g_t\|_2^2\\
&\le \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|_2^2\end{aligned}$$

Now we can use the smoothness of the losses to bound the right hand side, using the fact that the losses are bounded away from 0, and theorem 4.22, we have :

$$\forall t \in 1, \ldots, T, \ \|g_t\|_2^2 \leq 2M\ell_t(x_t)$$

Hence we get :

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{M}{\mu}\sum_{t=1}^{T}\frac{\ell_t(x_t)}{t}$$

Now we introduce $L_t = \sum_{s=1}^{t} \ell_s(x_s)$ and $L_t(u) = \sum_{s=1}^{t} \ell_s(u)$, we want to bound L_t, for that, we use the hypothesis that we made at the beginning that the minimum of each loss on V is equal to 0. Let x_t^* be such that $\ell_t(x_t^*) = 0$, we define the diameter of V as $D := \max_{x,y \in V} \|x - y\| < \infty$. We have :

$$|\ell_t(x_t)| = |\ell_t(x_t) - \ell_t(x_t^*)| \leq K\|x_t - x_t^*\| \leq DK$$

Where K is the Lipschitz constant of the losses. Then we have $L_t \leq DKt$ so $1 + L_t \leq 2DKt$ and $\frac{1}{t} \leq 2DK\frac{1}{1+L_t}$ Putting it all together, we get :

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{2DKM}{\mu}\sum_{t=1}^{T}\frac{\ell_t(x_t)}{1 + L_t}$$

We can now apply Lemma 4.13 with $f(x) = \frac{1}{x}$, $a_0 = 1$, $\forall 0 < t \leq T$, $a_t = \ell_t(x_t)$:

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{\ell_t(x_t)}{1 + L_t} \\
&= \sum_{t=1}^{T} a_t f\left(a_0 + \sum_{i=1}^{t} a_i\right) \\
&\leq \int_{a_0}^{\sum_{t=0}^{T} a_t} f(x)dx \\
&= \int_{1}^{1+L_T} \frac{1}{x}dx = \log(1 + L_T)
\end{aligned}$$

We now have :

$$L_T - L_T(u) \leq \frac{2DKM}{\mu}\log(1 + L_T)$$

In order to replace L_t with L^*, we will prove the following Lemma :

Lemma. *Let $0 \leq x, c, \alpha$ such that $x - c \leq \alpha\log(1 + x)$, then :*

$$x - c \leq \alpha^2 + \alpha\log(1 + c)$$

Proof. To prove this result, we will first prove a similar result with the square root and then compare the square root and the log.

We will prove that if $0 \leq x, c, \alpha$ are such that $x - c \leq \alpha\sqrt{x}$ and $\alpha^2 \leq x - c$, then $\sqrt{x} \leq \sqrt{c} + \alpha^2$

(In particular, this proves $x - c \leq \alpha\sqrt{x} \implies x - c \leq \alpha\sqrt{c} + \alpha^2$).

Since $\alpha^2 \leq x - c$ we have :

$$\begin{aligned} x - c + \alpha^2 &\leq 2\alpha\sqrt{x} \\ x + \alpha^2 - 2\alpha\sqrt{x} &\leq c \\ (\sqrt{x} - \alpha)^2 &\leq \sqrt{c} \\ \sqrt{x} - \alpha &\leq \sqrt{c} \\ \alpha\sqrt{x} &\leq \alpha^2 + \alpha\sqrt{c} \end{aligned}$$

Where we can take the square on the 4th line because everything is non-negative and the last line comes from multiplying by α. Now we compare the log and the square root. Let $f(x) = \log(1 + x)$, $g(x) = \sqrt{x}$. We have that

$$\forall x > 0, f'(x) - g'(x) = \frac{2\sqrt{x} - (1 + x)}{2\sqrt{x}(1 + x)} = \frac{-(1 - \sqrt{x})^2}{2\sqrt{x}(1 + x)} \leq 0$$

And $f(0) - g(0) = 0 \leq 0$ so $\forall x \geq 0, f(x) \leq g(x)$. We can finally prove our Lemma, let x, c, α be such that $x - c \leq \alpha \log(1 + x)$. Without loss of generality, we assume $c \leq x$ and $x - c \geq \alpha^2$ otherwise, there is nothing to prove. We have

$$x - c \leq \alpha f(x) \leq \alpha g(x) = \alpha\sqrt{x}$$

So by the previous computation :

$$\alpha\sqrt{x} - \alpha\sqrt{c} \leq \alpha^2$$

Now :

$$\begin{aligned} \alpha(\log(1 + x) - \log(1 + c)) &= \alpha(f(x) - f(c)) \\ &= \alpha \int_c^x f'(s)ds \\ &\leq \alpha \int_c^x g'(s)ds \\ &= \alpha(\sqrt{x} - \sqrt{c}) \\ &\leq \alpha^2 \end{aligned}$$

In the end :

$$x - c \leq \alpha \log(1 + x) \leq \alpha \log(1 + c) + \alpha^2$$

$\square$

Now we apply the previous Lemma with $x = L_T$, $c = L^*$, $\alpha = \frac{2DKM}{\mu}$:

$$L_T - L^* \leq \frac{2DKM}{\mu} \log(1 + L^*) + \left(\frac{2DKM}{\mu}\right)^2$$

Finally :

$$\sum_{t=1}^{T} \ell_t(x_t) - \ell_t(u) = L_T - L_t(u) \leq L_t - L^* \leq \frac{2DKM}{\mu} \log(1+L^*) + \left(\frac{2DKM}{\mu}\right)^2 = \mathcal{O}(\log(1+L^*))$$

$\square$

Chapter 6

Online Mirror Descent

Exercise 6.4

Let $A \in \mathbb{R}^{d\times d}$a positive definite matrix. Define $\|x\|_A^2 = x^T Ax$. Prove that $\frac{1}{2}\|x-y\|_A^2$ is the Bregman divergence $B_\varphi(x;y)$ associated with $\varphi(x) = \frac{1}{2}\|x\|_A^2$.

Solution : We have $\varphi(x) = \frac{1}{2}x^T Ax$ and $\nabla\varphi(x) = Ax$. Hence :

$$\begin{aligned}
B_{\varphi(x;y)} &= \varphi(x) - \varphi(y) - \langle \nabla\varphi(x), y - x\rangle \\
&= \frac{1}{2}x^T Ax - \frac{1}{2}y^T Ay - y^T A(x-y) \\
&= \frac{1}{2}\left(x^T Ax + y^T Ay - 2y^T Ax\right) \\
&= \frac{1}{2}\left((x-y)^T A(x-y)\right) \\
&= \frac{1}{2}\|x-y\|_A^2
\end{aligned}$$

$\square$