# Online Machine Learning Homework

Assignment 2

**Student:** Lei Feng, `please-help@each.other`
**Student ID:** 123456
**Lecturer:**

**Problem 1: The doubling trick**

You are given an online algorithm $\mathcal{A}$ that guarantees $Regret \leq T^p$ , for some $p \in (0, 1)$ , but it has parameters that must be chosen as a function of $T$. Using this algorithm as a black-box, we will construct an algorithm with a regret bound $\mathcal{O}(T^p)$ that holds simultaneously for all $T$. In particular, we will analyze the following transformation:

---

**for** epoch $m = 0, 1, 2, \ldots$ **do**
  Reset $\mathcal{A}$ with parameters chosen for $T = 2^m$
  **for** rounds $t = 2^m, \ldots, 2^{m+1} - 1$ **do**
    Run $\mathcal{A}$

---

Essentially, the algorithm initially guesses $T = 1$, and when it observes this guess was too low, it doubles it's initial guess and re-starts $\mathcal{A}$. Hence, this is called the "**doubling trick**".

To show the desired regret bound, consider any $T$, and

(a) Show that the *regret* on rounds 1 through $T$ is less than or equal to the *regret* on epochs $m = 0$ through the end of epoch $m_T = \lceil \log_2(T) \rceil$. Then, use the regret bound for $\mathcal{A}$ to bound the cumulative regret for these epochs.

(b) Simplify the bound from (a) to show that it is upper bounded by a constant times $T^p$. Hint: Use the fact that for $x \neq 1$,

$$\sum_{k=0}^{n} x^k = \frac{x^{n+1} - 1}{x - 1}$$

(a) According to the algorithm, within epoch $m$, $\mathcal{A}$ runs on rounds $T = 2^m, \ldots, 2^{m+1} - 1$. Hence, starting from epoch $m = 0$ through $m_T = \lceil \log_2(T) \rceil$, rounds $T = 2^0, 2^1, 2^1+1, 2^2, 2^2+1, \ldots, T$ are definitely executed, i.e., $\mathcal{A}$ runs on rounds 1 through $T$, and possibly goes on. Therefore,

$$\sum_{t=1}^{T} Regret(t) \leq \sum_{m=0}^{\lceil \log_2(T) \rceil} Regret(m)$$

Consider epoch m, the regret bound of $\mathcal{A}$ is computed for $T = 2^m$ times. Then

$$Regret(m) \leq T^p = 2^{mp}$$

Finally,

$$\sum Regret(m) \leq \sum_{m=0}^{\lceil \log_2(T) \rceil} 2^{mp}$$

(b) From (a), we established that

$$\sum Regret(m) \leq \sum_{m=0}^{m_T} 2^{mp}$$

where $m_T = \lceil \log_2(T) \rceil$ . Since $p \in (0,1)$, $2^p$ can be considered the term $x$ of the equation given in the hint. Thus,

$$\sum Regret(m) \leq \sum_{m=0}^{m_T} 2^{mp} = \frac{(2^p)^{m_T+1} - 1}{2^p - 1}.$$

As the number of epochs $m_T$ increases and approaches infinity, the $-1$ term in the numerator diminishes and becomes negligible, hence

$$\sum Regret(m) \leq \frac{2^{pm_T} \cdot 2^p}{2^p - 1}.$$

Finally, note that $m_T = \lceil \log_2(T) \rceil$, which implies that $\log_2 T \leq m_T < \log_2 T + 1$, we have

$$2^{pm_T} \leq (2^p)^{\log_2 T + 1} = T^p \cdot 2^p$$

Therefore,

$$\sum Regret(m) \leq \frac{2^{2p}}{2^p - 1} T^P$$

As demonstrated, the regret is upper bounded by a constant $C$ times $T^p$ where $C = \frac{2^{2p}}{2^p - 1}$ .

**Problem 2: Constructing a transformation to get stronger bounds**

In class we considered a one dimensional problem with linear loss functions $f_t(w) = g_t w$, where the adversary chooses $g_t \in [-1, 1]$ . The goal was low regret with respect to $\mathcal{W} = [-1, 1]$ . The Follow-The-Leader(FTL) algorithm did very badly when the adversary played $g_t$ according to the sequence $(0.5, 1, -1, 1, -1, \dots)$. We then showed that with an appropriate regularization term, the Follow-The-Regularized-Leader(FTRL) for linear functions achieves $Regret \leq \sqrt{2T}$ against the best fixed $w^* \in [-1, 1]$ (since $G = 1$ and $R = 1$).

However, in hindsight, one might not feel that competing with a *fixed* point is so great; after all a simple alternating strategy (playing $0, -1, 1, -1, 1, \dots$ ) would have achieved loss $\mathcal{O}(-T)$, while any fixed strategy has loss $\mathcal{O}(1)$. Show a transformation (using the FTRL algorithm as a subroutine) that gives a no-regret algorithm against a competitor set $\mathcal{W}'$ that includes this alternating strategy. Give the regret bound for this algorithm, and compare it to the regret bound achieved by applying FTRL directly to the problem.

Hint: Use a transformation that takes the original one-dimensional problem, and maps it into a two-dimensional online linear optimization problem. You will need to transform both the loss functions and the points played.

First, we extend the competitor set $\mathcal{W}$ to

$$\mathcal{W}' = \{(w_1, w_2) \mid w_1, w_2 \in [-1, 1]\},$$

then we define the new 2-dimensional loss function as:

$$\tilde{f}_t(w_1, w_2) = \begin{cases} g_t w_1 & \text{if } t \text{ is odd,} \\ g_t w_2 & \text{if } t \text{ is even.} \end{cases}$$

Then, at each round $t$, play:

$$w_t = \begin{cases} w_1 & \text{if } t \text{ is odd,} \\ w_2 & \text{if } t \text{ is even.} \end{cases}$$

Finally, use FTRL with a regularization term to ensure stability. The regularized loss for $T$ rounds is:

$$\tilde{F}_T(w_1, w_2) = \sum_{t=1}^{T} \tilde{f}_t(w_1, w_2) + R(w_1, w_2),$$

where $R(w_1, w_2)$ is a strongly convex regularization term, such as:

$$R(w_1, w_2) = \frac{1}{2}(w_1^2 + w_2^2).$$

The update rule for $(w_1, w_2)$ is:

$$(w_1^{t+1}, w_2^{t+1}) = arg \min_{(w_1, w_2) \in [-1,1]^2} \tilde{F}_t(w_1, w_2).$$

Now we have successfully transformed the problem from 1-dimensional to 2-dimensional. The regret bound however remains unchanged, $\mathcal{O}(\sqrt{T})$, and yet it can now compete against a competitor set $\mathcal{W}'$ that includes the alternating strategy.

**Problem 3: Convex functions and global lower bounds**

Recall that a function $f$ is *convex* if

$$f(\alpha x + (1-\alpha)x') \leq \alpha f(x) + (1-\alpha) f(x')$$

for any $\alpha \in [0, 1]$ and any $x$ and $x'$ in $f$'s domain. One of the key properties of convex functions is that a (sub)gradient of the function at a particular $w$ gives information about the global structure of the function. In particular:

(a) Prove that for a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$, for all $w$ and $w_0$ in the domain of $f$,

$$f(w) \geq f(w_0) + \nabla f(w_0)(w - w_0), \tag{1}$$

where $\nabla f(w_0)$ is the gradient of $f$ evaluated at $w_0$. That is, a first-order Taylor expansion of a convex function gives a lower bound on the function. Hint: Use the fact that

$$\nabla f(w) \cdot w' = \lim_{\delta \to 0} \frac{f(w + \delta w') - f(w)}{\delta}.$$

(b) Show that the previous condition is sufficient, that is, any function $f : \mathbb{R}^n \to \mathbb{R}$ such that Eq.(1) holds for all $w, w_0$ in the domain of $f$ is convex. Hint: Apply Eq.(1) twice at a carefully chosen point.

(c) Consider a convex $f$ and assume a $w^* \in arg\min_w f(w)$ exists. (Aside: often we write $w^* \in arg\min_w f(w)$, but this is sloppy, because the $arg\min$ need not be unique. This sloppiness is usually fine, because we don't care which $w^* \in arg\min_w f(w)$ we get. Technically, we define ( $arg\min_{w\in\mathcal{W}} f(w) = \{w^* \in \mathcal{W} \mid f(w^*) \leq f(w), \forall w \in \mathcal{W}\}$. ) Show that by evaluating $f$ and computing its gradient at any point $w$, we can find a half-space that contains $w^*$ (and hence a half-space that does not contain $w^*$). Recall that a half-space is a set of points $\{w \mid a \cdot w \geq b\}$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

(d) Consider a convex $f$ in one dimension, defined on $[0, D]$, such that there exists a $w^* \in arg\min_{w\in[0,D]} f(w)$. Show that we can find a $w'$ such that $|w^* - w'| \leq \epsilon$ by making only $\lfloor \log_2 \frac{D}{\epsilon} \rfloor$ queries to an oracle that computes $\nabla f(w)$.

(e) Suppose $\vec{0} \in \mathbb{R}^n$ is a subgradient of a convex function $f : \mathcal{W} \to \mathbb{R}^n$ at $w^*$ with $f(w^*)$ finite. Show that $w^* \in arg\min_{w\in\mathcal{W}} f(w)$.

(a) *Proof.* By definition of convexity, for any $w, w_0 \in \mathrm{Dom}(f)$ and $\alpha \in [0, 1]$,

$$f(\alpha w + (1-\alpha)w_0) \leq \alpha f(w) + (1-\alpha)f(w_0).$$

Let $w' = w - w_0$ and substitute $\alpha$ with $\delta \in [0, 1]$ to consider the point $w_\delta = w_0 + \delta w'$:

$$\begin{aligned} f(w_\delta) &= f(\delta w' + w_0) \\ &= f(\delta(w - w_0) + w_0) \\ &= f(\delta w + (1 - \delta w_0)) \leq \delta f(w) + (1-\delta)f(w_0). \end{aligned}$$

Rearranging terms, dividing by $\delta$, and taking the limit as $\delta \to 0^+$, we have

$$\lim_{\delta\to 0^+} \frac{f(w_\delta) - f(w_0)}{\delta} \leq f(w) - f(w_0).$$

By the definition of the gradient, the left-hand side equals $\nabla f(w_0) \cdot w'$,

$$\nabla f(w_0) \cdot (w - w_0) \leq f(w) - f(w_0).$$

Thus we have

$$f(w) \geq f(w_0) + \nabla f(w_0) \cdot (w - w_0).$$

□

(b) Assume that $f$ satisfies Eq. (1) for all $w, w_0 \in \mathrm{dom}(f)$. Let $w, w' \in \mathrm{dom}(f)$ and $\alpha \in [0, 1]$. Define $w_\alpha = \alpha w + (1-\alpha)w'$. Applying Eq. (1) at $w_0 = w$ and $w_0 = w'$, we have:

$$f(w_\alpha) \geq f(w) + \nabla f(w) \cdot (w_\alpha - w),$$

$$f(w_\alpha) \geq f(w') + \nabla f(w') \cdot (w_\alpha - w').$$

Taking a convex combination with weights $\alpha$ and $1 - \alpha$, we get

$$f(w_\alpha) \geq \alpha f(w) + (1-\alpha)f(w'),$$

thus showing $f$ is convex.

(c) Let $f$ be convex and let $w^* \in \arg\min_w f(w)$. By Eq. (1), for any $w \in \text{dom}(f)$:

$$f(w) \geq f(w^*) + \nabla f(w^*) \cdot (w - w^*).$$

Since $f(w^*) \leq f(w)$, it follows that:

$$\nabla f(w^*) \cdot (w - w^*) \geq 0.$$

Thus, the half-space $\{w \mid \nabla f(w^*) \cdot (w - w^*) \geq 0\}$ contains $w^*$.

(d) Assume $f$ is convex on $[0, D]$, and let $w^* \in \arg\min_{w \in [0,D]} f(w)$. At each step:

- Query the gradient $\nabla f(w)$ at the midpoint $w_m = \frac{a+b}{2}$ of the interval $[a, b]$.
- If $\nabla f(w_m) > 0$, set $b = w_m$; if $\nabla f(w_m) < 0$, set $a = w_m$; if $\nabla f(w_m) = 0$, stop and return $w^* = w_m$.

After $k$ queries, the interval length is $\frac{D}{2^k}$. To ensure $|w^* - w'| \leq \epsilon$, we need:

$$\frac{D}{2^k} \leq \epsilon \implies k \geq \log_2 \frac{D}{\epsilon}.$$

Thus, $\lfloor \log_2 \frac{D}{\epsilon} \rfloor$ queries suffice.

(e) Let $\vec{0} \in \partial f(w^*)$, the subdifferential of $f$ at $w^*$. By the definition of subgradients, for all $w \in \text{Dom}(f)$:

$$f(w) \geq f(w^*) + \vec{0} \cdot (w - w*).$$

This simplifies to:

$$f(w) \geq f(w^*).$$

Thus, $w^* \in \arg\min_w f(w)$.

**Problem 4: Convex sets and randomization**

A set $C$ is convex if for any $w_1, w_2 \in C$, and any $\alpha \in [0, 1]$, we have a $\alpha w_1 + (1 - \alpha) w_2 \in C$.

(a) Let $\mathcal{W} \subseteq \mathbb{R}^n$ be a convex set, with $w_1, \ldots, w_k \in \mathcal{W}$, and let $\theta_1, \ldots, \theta_k \in \mathbb{R}$ that satisfy $\theta_i \geq 0$ and $\sum_{i=1}^k \theta_i = 1$. Show that $\bar{w} = \sum_{i=1}^k \theta_i x_i$ is also in $\mathcal{W}$. We say that $\bar{w}$ is a **convex combination** of the $w_i$.

(b) Now, let $w_1, \ldots, w_k \in \mathbb{R}^n$ be arbitrary points, and let

$$\Delta^k = \{\theta \in \mathbb{R}^k \mid \theta_i \geq 0, \sum_{i=1}^k \theta_i = 1\}$$

be the $k$-dimensional probability simplex (the set of probability distributions on $k$ items). Show that the convex hull of the $w_i$,

$$\text{conv}(w_i, \ldots, w_k) = \{\theta \cdot w \mid \theta \in \Delta^k\}$$

is in fact a convex set.

(c) Let $w_1, \ldots, w_k \in \mathbb{R}^n$ be arbitrary points, let $\mathcal{W} = \text{conv}(w_i, \ldots, w_k)$, and let $f(w = g \cdot w)$ be a linear loss function on $\mathcal{W}$. Show that for any $w \in \mathcal{W}$, there exists a probability distribution such that choosing a $w_i$ according to the distribution and then playing the chosen $w_i$ against $f$ produces the same expected loss as just playing $w$. Conversely, show that for any probability distribution on $w_1, \ldots, w_k$ there exists a $w \in \mathcal{W}$ that gets the same expected regret. When might it be preferable to represent such a strategy as a distribution $\theta \in \Delta^k$, and when might it be preferable to represent such a strategy as a point $w \in \mathcal{W}$ ? (Hint: consider $n$ and $k$).

(a) By the definition of Convex sets,

$$\theta_1 w_1 + \theta_2 w_2 \in \mathcal{W},$$

where $\theta_1 + \theta_2 = 1, w_1, w_2 \in \mathcal{W}$.

Note that $\frac{\theta_1}{\theta_1+\theta_2} + \frac{\theta_2}{\theta_1+\theta_2} = 1$. Therefore, it holds that

$$\hat{w_2} = \frac{\theta_1}{\theta_1 + \theta_2} w_1 + \frac{\theta_2}{\theta_1 + \theta_2} w_2 \in \mathcal{W}.$$

Now that it is proven that $\hat{w_2} \in \mathcal{W}$, we apply the definition of convex sets again, and then we get the combination $(\theta_1 + \theta_2)\hat{w_2} + (1 - \theta_1 - \theta_2)w_3 \in \mathcal{W}$.

Equivalently,

$$\begin{aligned}(\theta_1 + \theta_2)\hat{w_2} + \theta_3 w_3 &= (\theta_1 + \theta_2)\big(\frac{\theta_1}{\theta_1 + \theta_2} w_1 + \frac{\theta_2}{\theta_1 + \theta_2} w_2\big) + \theta_3 w_3 \\ &= \theta_1 w_1 + \theta_2 w_2 + \theta_3 w_3 \in \mathcal{W}\end{aligned}$$

Hence, we proved that the **convex combination** of $w_1, w_2, w_3$ is also in $\mathcal{W}$.

We can now apply this property **inductively**, for example, by defining

$$\hat{w_3} = \frac{\theta_1}{\theta_1 + \theta_2 + \theta_3} w_1 + \frac{\theta_2}{\theta_1 + \theta_2 + \theta_3} w_2 + \frac{\theta_3}{\theta_1 + \theta_2 + \theta_3} w_3$$

and assert its presence in $\mathcal{W}$ by pointing out that its coefficients add up to 1. Then $\hat{w_3}$'s presence can be proven since we have already established that the convex combination of $w_1, w_2, w_3$ is also in $\mathcal{W}$.

(b) Let $w_1 = \theta_1 \cdot w$ and $w_2 = \theta_2 \cdot w$, where $\theta_1, \theta_2 \in \Delta^k$.

$$\alpha w_1 + (1 - \alpha)w_2 = \alpha(\theta_1 \cdot w) + (1 - \alpha)(\theta_2 \cdot w) = (\alpha\theta_1 + (1 - \alpha)\theta_2) \cdot w$$

According to the definition of $\Delta^k$, we have $\alpha\theta_1 + (1-\alpha)\theta_2 \in \Delta^k$ . Therefore, $\alpha w_1 + (1-\alpha)w_2$ is in $\text{conv}(w_i, ..., w_k)$.

(c) Let $\mathcal{W} = \text{conv}(w_1, \ldots, w_k)$, and let $f(w) = g \cdot w$ be a linear loss function. For any $w \in \mathcal{W}$, $w = \sum_{i=1}^{k} \theta_i w_i$ for some $\theta \in \Delta^k$. The expected loss when choosing $w_i$ with probability $\theta_i$ is:

$$\begin{aligned}
\mathbb{E}[f(w_i)] &= \sum_{i=1}^{k} \theta_i f(w_i) \\
&= \sum_{i=1}^{k} \theta_i (g \cdot w_i) \\
&= g \cdot \left( \sum_{i=1}^{k} \theta_i w_i \right) \\
&= g \cdot w = f(w).
\end{aligned}$$

Thus, playing $w \in \mathcal{W}$ is equivalent to choosing $w_i$ according to $\theta$.

Conversely, for any $\theta \in \Delta^k$, let $w = \sum_{i=1}^{k} \theta_i w_i$. Then the loss from playing $w$ is:

$$f(w) = g \cdot w = \sum_{i=1}^{k} \theta_i (g \cdot w_i) = \sum_{i=1}^{k} \theta_i f(w_i).$$

Thus, playing $w \in \mathcal{W}$ achieves the same expected loss as the distribution $\theta$.

Last but not least, use $\theta \in \Delta^k$ when $k \ll n$, vice versa.