Machine Learning Theory (CSC 431/531) - Lectures 19 and 20

Nishant Mehta

1 Online Convex Optimization and Online Gradient Descent

One of the most fundamental settings in online learning is Online Convex Optimization (OCO). Let us see how this game is defined. Let the action space be a closed convex set $V \subseteq \mathbb{R}^d$. The game protocol is given below.

Algorithm 1: ONLINE CONVEX OPTIMIZATION

```
for $t = 1 \to T$ do
    Learner plays $w_t \in V \subseteq \mathbb{R}^d$
    Nature plays convex loss function $\ell_t \colon V \to \mathbb{R}$
    Learner suffers loss $\ell_t(w_t)$
end
```

Although the losses ℓ_t need not be differentiable, we can introduce the main ideas even in the differentiable case. Therefore, for simplicity, we begin with the assumption that the losses are differentiable. We consider the more general case of potentially non-differentiable losses later in this lecture; this case is important as even the absolute loss is not differentiable.

The first algorithm we will study for OCO is called (Projected) Online Gradient Descent (OGD). If you have seen stochastic gradient descent before, the algorithm should look very familiar. Before presenting the algorithm, we first introduce the notion of Euclidean projection.

For a nonempty, closed convex subset $A \subseteq \mathbb{R}^d$, the Euclidean projection of y onto A is

$$\Pi_A(y) = \operatorname*{arg\,min}_{x \in A} \|x - y\|_2.$$

Note that A is closed to ensure the minimizer belongs to A, nonempty for obvious reasons, and convex because this ensures that the minimizer is unique (owing to the strong convexity of the equivalent objective $w \mapsto \|y - w\|_2^2$).

Algorithm 2: (PROJECTED) ONLINE GRADIENT DESCENT

```
Select any $w_1 \in V$
for $t = 1 \to T$ do
    Play $w_t$
    Observe loss function $\ell_t$ and suffer loss $\ell_t(w_t)$
    $w_{t+1} \leftarrow \Pi_V(w_t - \eta g_t)$ for $g_t = \nabla \ell_t(w_t)$
end
```

Before analyzing OGD, we need a couple of tools from convex analysis. A fundamental property of convex functions is that they can be lower bounded by their first-order Taylor approximation.

Theorem 1. *Let $V \subseteq \mathbb{R}^d$ be convex. Let $f : V \to \mathbb{R}$ be convex and $x \in V$. If f is differentiable at x, then*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \textit{for all } y \in V.$$

It turns out that a variation of the above result can still be established even if f is not differentiable at x; we will see this extension later in this lecture.

The second tool that we need is that projections can only reduce Euclidean distance.[1]

Proposition 1. *Let $V \subseteq \mathbb{R}^d$ be closed and convex. For any $w \in \mathbb{R}^d$ and $u \in V$,*

$$\|\Pi_V(w) - u\|_2 \leq \|w - u\|_2.$$

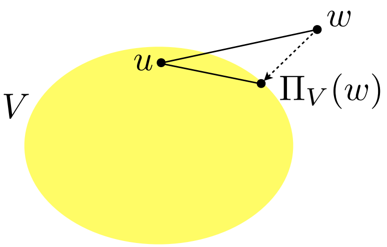

Figure 1: Euclidean projection of w onto a convex set V can only reduce distance to u.

We now have the tools we need to prove a regret bound for OGD.

Theorem 2. *Let $V \subset \mathbb{R}^d$ be a closed convex set with diameter $\max_{x,y \in V} \|x - y\|_2 \leq D$. Let $\ell_1, \ell_2, \ldots, \ell_T$ be a sequence of convex loss functions $\ell_t : V \to \mathbb{R}$ that are differentiable on V. For any $w_1 \in V$ and learning rate $\eta > 0$, for any comparator $u \in V$, we have*

$$\sum_{t=1}^{T} (\ell_t(w_t) - \ell_t(u)) \leq \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_2^2.$$

Moreover, letting $G := \max_{t \in [T]} \|g_t\|_2 \leq G$ and setting $\eta = \frac{D}{G\sqrt{T}}$, the bound becomes $DG\sqrt{T}$.

Proof. The second statement in the theorem is immediate from the suggested setting of η.

We now prove the first statement. The proof is in four steps. The first step is just like our analysis of the Exponentially Weighted Average Forecaster for convex losses.

First step: Linearization

Recall that $g_t = \nabla \ell_t(w_t)$. Theorem 1 implies that

$$\ell_t(w_t) - \ell_t(u) \leq \langle g_t, w_t - u \rangle. \tag{1}$$

[1] For a proof, see the proof of Proposition 1 of http://web.uvic.ca/~nmehta/online_learning_spring2023/lecture3.pdf.

Second step: Using recurrence via warm-up analysis

It will be useful to define the intermediate iterate $\widetilde{w}_{t+1} = w_t - \eta g_t$, which is just w_{t+1} if we did not project back onto the feasible set V. Now, note that we may rewrite the 2η-weighted linearized regret as

$$\begin{aligned} 2\eta\langle g_t, w_t - u\rangle &= \|w_t - u\|_2^2 - \|w_t - u - \eta g_t\|_2^2 + \eta^2\|g_t\|_2^2 \\ &= \|w_t - u\|_2^2 - \|\widetilde{w}_{t+1} - u\|_2^2 + \eta^2\|g_t\|_2^2, \end{aligned}$$

so that

$$\langle g_t, w_t - u\rangle = \frac{1}{2\eta}\|w_t - u\|_2^2 - \frac{1}{2\eta}\|\widetilde{w}_{t+1} - u\|_2^2 + \frac{\eta}{2}\|g_t\|_2^2. \tag{2}$$

Third step: Relating $\widetilde{w}_{t+1}$ to actual iterate w_{t+1}

It would be nice if the expression (2) had w_{t+1} in place of $\widetilde{w}_{t+1}$ as we could then sum this expression over t to get a telescoping series. We can make this swap with the help of Proposition 1 since $w_{t+1} = \Pi(\widetilde{w}_{t+1})$ and hence $\|w_{t+1} - u\|_2^2 \leq \|\widetilde{w}_{t+1} - u\|_2^2$, which gives

$$\langle g_t, w_t - u\rangle \leq \frac{1}{2\eta}\|w_t - u\|_2^2 - \frac{1}{2\eta}\|w_{t+1} - u\|_2^2 + \frac{\eta}{2}\|g_t\|_2^2.$$

Fourth step: Summing over t and telescoping

Combining this inequality with (1) and summing over t gives that twice the regret is bounded as

$$\begin{aligned} &2\sum_{t=1}^{T}(\ell_t(w_t) - \ell_t(u)) \\ &\leq \sum_{t=1}^{T}\left(\frac{1}{\eta}\|w_t - u\|_2^2 - \frac{1}{\eta}\|w_{t+1} - u\|_2^2 + \eta\|g_t\|_2^2\right) \\ &= \frac{1}{\eta}\|w_1 - u\|_2^2 - \frac{1}{\eta}\|w_{T+1} - u\|_2^2 + \eta\sum_{t=1}^{T}\|g_t\|_2^2 \\ &\leq \frac{1}{\eta}\|w_1 - u\|_2^2 + \eta\sum_{t=1}^{T}\|g_t\|_2^2 \\ &\leq \frac{D^2}{\eta} + \eta\sum_{t=1}^{T}\|g_t\|_2^2. \end{aligned}$$

This proves the first statement of the theorem. □

2 Online Subgradient Descent

Projected Online Gradient Descent relies upon the ability to receive a gradient $g_t = \nabla\ell_t(w_t)$ in each round. However, there are many popular loss functions that fail to be differentiable at some points in V. For example:

- Taking $V = \mathbb{R}$ and $y_t \in \mathbb{R}$, the absolute loss $\ell_t(w_t) = |w_t - y_t|$ is not differentiable at $w_t = y_t$;

- Taking $V = \mathbb{R}^d$, $x_t \in \mathbb{R}^d$, and $y_t \in \{-1, +1\}$, the hinge loss $\ell_t(w) = \max\{0, 1 - y_t\langle w, x_t\rangle\}$ is not differentiable whenever $y_t\langle w, x_t\rangle = 1$.

It turns out that for both these loss functions as well as many others, we can still use OGD after making a small but important change. Recall that the whole reason we required differentiability of the losses was so that we could upper bound the regret by its linear approximation via Theorem 1 (see the first step of the proof of Theorem 2).

It would be very nice if, even at a point $x \in V$ where a function is not differentiable, we had available some $g \in \mathbb{R}^d$ that is *like* a gradient in the sense that

$$f(y) \geq f(x) + \langle g, y - x\rangle \quad \text{for all } y \in V.$$

There indeed is such a concept.

Definition 1 (Subgradient)**.** Let $V \subseteq \mathbb{R}^d$ be convex. Let $f \colon V \to \mathbb{R}$ be convex. A vector $g \in \mathbb{R}^d$ is a *subgradient* of f at $x \in V$ if

$$f(y) \geq f(x) + \langle g, y - x\rangle \quad \text{for all } y \in V. \tag{3}$$

In general, there can be a set of subgradients of f at x, and this set is called the *subdifferential* of f at x, denoted as $\partial f(x)$. When f is differentiable at x, then there is only one subgradient of f at x, and this subgradient is the gradient $\nabla f(x)$. Conversely, if there is only one subgradient of f at x, then *(i)* this subgradient must be $\nabla f(x)$ and *(ii)* f is differentiable at x. Because we assume that f is real-valued, it turns out that $\partial f(x)$ is always non-empty, which means that f is *subdifferentiable* at x.

Let us see an example.

Example 1 (Subdifferential of absolute loss)**.** Setting $y = 0$, the absolute loss is $\ell(w) = |w|$. Clearly, ℓ is differentiable whenever $w \neq 0$. When $w < 0$, we have $\partial\ell(w) = \{-1\}$, and when $w > 0$, we have $\partial\ell(w) = \{+1\}$. Let us investigate what happens when $w = 0$. The condition (3) becomes

$$\ell(w) \geq \ell(0) + g \cdot (w - 0) \quad \text{for all } w \in \mathbb{R},$$

or

$$|w| \geq g \cdot w \quad \text{for all } w \in \mathbb{R},$$

which is satisfied if and only if $g \in [-1, 1]$. Therefore, $\partial\ell(0) = [-1, 1]$. To recap, we have

$$\partial\ell(w) = \begin{cases} -1 & \text{if } w < 0; \\ [-1, 1] & \text{if } w = 0; \\ 1 & \text{if } w > 0. \end{cases}$$

How should we modify OGD when we are only guaranteed to have subgradients rather than gradients? Well, we just use any subgradient in each round. This modification of OGD (which is often still called OGD) is (Projected) Online Subgradient Descent.

Algorithm 3: (Projected) Online Subgradient Descent

```
Select any $w_1 \in V$
for $t = 1 \to T$ do
    Play $w_t$
    Observe loss function $\ell_t$ and suffer loss $\ell_t(w_t)$
    $w_{t+1} \leftarrow \Pi_V(w_t - \eta g_t)$ for any $g_t \in \partial \ell_t(w_t)$
end
```

Notice that the only step of the proof of Theorem 2 that needs to be modified is the first one, and as already mentioned above, the step still goes through if we use a subgradient $g_t \in \partial \ell_t(w_t)$ in round t. The following regret bound for Online Subgradient Descent is immediate.

Theorem 3. *Let $V \subset \mathbb{R}^d$ be a closed convex set with diameter $\max_{x,y \in V} \|x - y\|_2 \leq D$. Let $\ell_1, \ell_2, \ldots, \ell_T$ be a sequence of convex loss functions $\ell_t \colon V \to \mathbb{R}$ that are subdifferentiable on V. Let $(g_t)_{t \geq 1}$ be any sequence of subgradients $g_t \in \partial \ell_t(w_t)$. For any $w_1 \in V$ and any learning rate $\eta > 0$, for any comparator $u \in V$, we have*

$$\sum_{t=1}^{T} (\ell_t(w_t) - \ell_t(u)) \leq \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_2^2.$$