**Advanced Topics in ML and AGT** November 27, 2017

# Lecture 6

*Lecturer: Yishay Mansour* *Scribes: Aviv Rosenberg, Daniel Nurieli, Elias Hanna*

## 1 Online Mirror Descent (with estimated gradients)

Recall the Online Mirror Descent (OMD) algorithm we described in Lecture 4. Now suppose that instead of setting $z_t$ to be a sub-gradient of $f_t$ at $w_t$, we shall set $z_t$ to be a random vector with $E[z_t] \in \partial f_t(w_t)$.

**Algorithm 1** Online Mirror Descent with estimated gradients

set $y_1$ such that $\nabla R(y_1) = 0$
set $w_1 = \arg\min_{w \in S} B_R(w||y_1)$
**for** $t = 1, ..., T$ **do**
    play $w_t$
    get $z_t$ such that $E[z_t|z_{t-1}, ..., z_1] \in \partial f_t(w_t)$
    set $\nabla R(y_{t+1}) = \nabla R(y_t) - \eta z_t$
    set $w_{t+1} = \arg\min_{w \in S} B_R(w||y_{t+1})$

The following theorem tells us how to extend previous regret bounds we derived for OMD to the case of estimated sub-gradients.

**Theorem 1** *Suppose that the Online Mirror Descent with Estimated Gradients is run on a sequence of loss functions $f_1, ..., f_T$. Suppose that the estimated sub-gradients are chosen such that*

$$\sum_{t=1}^{T} (w_t - u)^T z_t \leq B(u) + \sum_{t=1}^{T} ||z_t||_t^2$$

*where $B$ is some function and for all round $t$ the norm $||\cdot||_t$ may depend on $w_t$. Then,*

$$E[regret] = E\left[\sum_{t=1}^{T} f_t(w_t) - f_t(u)\right] \leq B(u) + \sum_{t=1}^{T} E[||z_t||_t^2]$$

*where expectation is with respect to the randomness in choosing $z_1, ..., z_t$.*

**Proof:** Taking expectation of both sides of the first inequality with respect to the randomness in choosing $z_t$ we obtain that

$$\sum_{t=1}^{T} E\left[(w_t - u)^T z_t\right] \leq B(u) + \sum_{t=1}^{T} E[||z_t||_t^2]$$

At each round, let $v_t = E[zt|z_{t1}, ..., z_1]$. By the law of total probability we obtain that

$$\sum_{t=1}^{T} E\left[(w_t - u)^T z_t\right] = \sum_{t=1}^{T} E\left[(w_t - u)^T v_t\right]$$

Since we assume that $v_t \in \partial f_t(w_t)$ we know that

$$(w_t - u)^T v_t \geq f_t(w_t) - f_t(u)$$

Combining all the above we conclude our proof. ■

The above theorem tells us that as long as we can find $z_1, ..., z_T$ which on one hand are unbiased estimators of sub-gradients and on the other hand has bounded norms, we can still obtain a valid regret bound.

## 2 The Multi-armed Bandit Problem

In the multi-armed bandit problem, there are $d$ arms, and on each online round the learner should choose one of the arms, denoted $p_t$, where the chosen arm can be a random variable. Then, it receives a cost of choosing this arm, $y_t[p_t] \in [0, 1]$. The vector $y_t \in [0, 1]^d$ associates a cost for each of the arms, but the learner only gets to see the cost of the arm it pulls.

This problem is similar to prediction with expert advice. The only difference is that the learner does not get to see the cost of experts he didn't choose. The goal of the learner is to have low regret for not always pulling the best arm,

$$regret = E\left[\sum_{t=1}^{T} y_t[p_t]\right] - \min_{1 \leq i \leq d} \sum_{t=1}^{T} y_t[i]$$

where the expectation is over the learners own randomness.

This problem nicely captures the exploration-exploitation trade off. On one hand, we would like to pull the arm which, based on previous rounds, we believe has the lowest cost. On the other hand, maybe it is better to explore the arms and find another arm with a smaller cost.

To approach the multi-armed bandit problem we use the OMD with estimated gradients method derived in the previous section. As in the Weighted Majority algorithm for prediction with expert advice, we let $S$ be the probability simplex and the loss functions be $f_t(w) = w^T y_t$. The learner picks an arm according to $\Pr[p_t = i] = w_t[i]$ and therefore $f_t(w_t)$ is the expected cost of the chosen arm. The gradient of the loss function is $y_t$. However, we do not know the value of all elements of $y_t$, we only get to see the value $y_t[p_t]$. To estimate the gradient, we define the random vector $z_t$ as follows:

$$z_t[j] = \begin{cases} \frac{y_t[j]}{w_t[j]}, & p_t = j \\ 0, & p_t \neq j \end{cases}$$

We indeed have that $z_t$ is an unbiased estimate of the gradient because

$$E[z_t[j]] = \Pr[p_t = j]\frac{y_t[j]}{w_t[j]} + 0 = w_t[j]\frac{y_t[j]}{w_t[j]} = y_t[j]$$

We update $w_t$ using the update rule of the normalized EG algorithm. The resulting algorithm is given below.

**Algorithm 2** Multi-armed Bandit Algorithm (EXP3)

parameter: $\eta > 0$
set $w_1 = (1/d, ..., 1/d)$
**for** $t = 1, ..., T$ **do**
    choose $p_t \sim w_t$ and play $p_t$
    observe $y_t[p_t]$
    update $\hat{w}_t[p_t] = w_t[p_t] e^{-\eta \frac{y_t[p_t]}{w_t[p_t]}}$
    for $i \neq p_t$ update $\hat{w}_t[i] = w_t[i]$
    update $w_{t+1}[i] = \frac{\hat{w}_t[i]}{\sum_{j=1}^{d} \hat{w}_t[j]}$

**Theorem 2** *The multi-armed bandit algorithm enjoys the bound*

$$E\left[\sum_{t=1}^{T} y_t[p_t]\right] \leq \min_{1 \leq i \leq d} \sum_{t=1}^{T} y_t[i] + \frac{\log d}{\eta} + \eta d T$$

*In particular, setting* $\eta = \sqrt{\log d / dT}$ *we obtain the regret bound of* $2\sqrt{dT \log d}$.

**Proof:** For the Weighted Majority algorithm we have seen that

$$\sum_{t=1}^{T} (w_t - u)^T z_t \leq \frac{\log d}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{d} w_t[i] z_t[i]^2$$

So our previous OMD with estimated gradients theorem gives us that

$$E\left[\sum_{t=1}^{T} f_t(w_t) - f_t(u)\right] \leq \frac{\log d}{\eta} + \eta \sum_{t=1}^{T} E\left[\sum_{i=1}^{d} w_t[i] z_t[i]^2\right]$$

The last term can be bounded as follows:

$$E\left[\sum_{i=1}^{d} w_t[i] z_t[i]^2 | z_{t-1}, ..., z_1\right] = \sum_{j=1}^{d} \Pr[p_t = j] \sum_{i=1}^{d} w_t[i] z_t[i]^2$$

$$= \sum_{j=1}^{d} w_t[j] w_t[j] \frac{y_t[j]^2}{w_t[j]^2} = \sum_{j=1}^{d} y_t[j]^2 \leq d$$

which finishes the proof. ■

Comparing the above bound to the bound we derived for the Weighted Majority algorithm we observe an additional factor of $d$, which intuitively stems from the fact that here we only receive $1/d$ of the feedback the Weighted Majority algorithm receives. It is possible to rigorously show that the dependence on $d$ is unavoidable and that the bound we derived is essentially tight. (We will do it in one of the following lectures when proving lower bound of $\Omega(\sqrt{Td})$ for bandits problem. Also, there is another algorithm, INF, that achieves this lower abound and has regret $O(\sqrt{Td})$.)

# 3 Bandit Convex Optimization

In this section we consider the general online convex optimization problem, where we only have a black-box access to the loss functions, and thus cannot calculate sub-gradients directly. Like in our previous setting we will pick $w_t$, but only observe $f_t(w_t)$ and not all of $f_t$. So our main challenge will be to construct an unbiased estimator for our sub-gradient while we only observe one point.
An idea for solving our issue: $f'(x) \approx \frac{f(x+\delta)-f(x-\delta)}{2\delta} = E_{b\sim\{+1,-1\}}[\frac{f(x+b\delta)}{\delta}b]$.
Therefore if we sample $b$ and compute $\frac{f(x+b\delta)}{\delta}b$ we get an unbiased estimator for our sub-gradient. Now let's construct it formally.

**Definition 3** $U_b = \{v \in \mathbb{R}^d : ||v|| \leq 1\}$

**Definition 4** $U_{sp} = \{u \in \mathbb{R}^d : ||u|| = 1\}$

**Definition 5** *Given $\delta > 0$ we define a smoothed version of $f$ as follows:*

$$\hat{f}(w) = E_{v\sim U_b}[f(w + \delta v)]$$

As we will show below, the advantage of $\hat{f}$ is that it is differentiable and we can estimate its gradient using a single oracle call to $f$. But, before that, we show that $\hat{f}$ is similar to $f$.

**Lemma 6** *If $f$ is $L$-Lipschitz then $|\hat{f}(w) - f(w)| \leq \delta L$.*

**Proof:** By the Lipschitzness $|f(w + \delta u) - f(w)| \leq L||\delta u|| \leq \delta L$ (Since $||u|| \leq 1$). ■

We next show that the gradient of $\hat{f}$ can be estimated using a single oracle call to $f$.

**Lemma 7** $E_{u\sim U_{sp}}[\frac{d}{\delta}f(w + \delta u)u] = \nabla\hat{f}(w)$

**Proof:** For $d = 1$ let $F$ be the anti-derivative of $f$, namely, $F'(w) = f(w)$. By the fundamental theorem of calculus we have

$$\int_{-\delta}^{+\delta} f(w + t)dt = F(w + \delta) - F(w - \delta)$$

Note that in the 1-dimensional case we have that $v$ is distributed uniformly over $[1, 1]$ and

$$\hat{f}(w) = E_{v\sim[-1,1]}[(f(w + \delta v)] = \frac{\int_{-\delta}^{+\delta} f(w + t)dt}{2\delta} = \frac{F(w + \delta) - F(w - \delta)}{2\delta}$$

It follows that

$$\hat{f'(w)} = \frac{F'(w + \delta) - F'(w - \delta)}{2\delta} = \frac{f(w + \delta) - f(w - \delta)}{2\delta} = E_{u\sim\{-1,+1\}}[\frac{1}{\delta}f(w + \delta u)u]$$

For the case of $d > 1$ we will use Stokes' Theorem:

$$\nabla \int_{U_b} f(w + \delta v)dv = \int_{U_{sp}} f(w + \delta u)udu$$

Let $\texttt{Vol-b-}\delta$ be the volume of a ball of radius $\delta$ in $d$-dimensions and $\texttt{Vol-sp-}\delta$ the area of the sphere (the boundary of the ball). Notice that

$$\hat{f}(w) = E_{v\sim U_b}[(F(w+\delta v)] = \frac{1}{\texttt{Vol-b-}\delta}\int_{U_b^\delta} f(w+v)dv$$

$$E_{u\sim U_{sp}}[f(w+\delta u)u] = \frac{1}{\texttt{Vol-sp-}\delta}\int_{U_{sp}^\delta} f(w+u)\frac{u}{||u||}du$$

And now from Stokes' it follows that

$$\nabla\hat{f}(w) = \frac{1}{\texttt{Vol-b-}\delta}\nabla\int_{U_b^\delta} f(w+v)dv = \frac{1}{\texttt{Vol-b-}\delta}\int_{U_{sp}^\delta} f(w+u)\frac{u}{||u||}du =$$

$$\frac{\texttt{Vol-sp-}\delta}{\texttt{Vol-b-}\delta}E_{u\sim U_{sp}}[f(w+\delta u)u] = \frac{d}{\delta}E_{u\sim U_{sp}}[f(w+\delta u)u]$$

Where the final equality is due to the ratio of the volume of the ball and the sphere. ■

Now we can consider the resulting algorithm.

**Algorithm 3** Bandit Online Gradient Descent

parameter: $\delta > 0$, $\eta > 0$, convex $S \subseteq \mathbb{R}^d$
set $\theta_1 = 0$
**for** $t = 1, ..., T$ **do**
    pick $v_t \sim U_{sp}$
    use $w_t + \delta v_t$
    observe $f_t(w_t + \delta v_t)$
    set $z_t = \frac{d}{\delta} f_t(w_t + \delta v_t) v_t$
    update $\theta_{t+1} = \theta_t - \eta z_t$
    update $w_{t+1} = \arg\min_{w\in S} ||w - \theta_{t+1}||_2$

Let us now analyze the regret of this algorithm.

**Theorem 8** *Consider running the Bandit Online Gradient Descent algorithm on a sequence $f_1, ..., f_T$ of L-Lipschitz functions. Let $S$ be a convex set and define $B = \max_{u\in S} ||u||$ and $F = \max_{u\in S, t\in[T]} f_t(u)$. Then, for all $u \in S$ we have*

$$E[\sum_{t=1}^{T} f_t(w_t + \delta v_t) - f_t(u)] \leq 3L\delta T + \frac{B^2}{2\eta} + \eta T d^2 (F/\delta + L)^2$$

*In particular, setting $\eta = \frac{B}{d(F/\delta+L)\sqrt{2T}}$ and $\delta = \sqrt{\frac{BdF}{3L}}T^{-1/4}$ we obtain that the regret is bounded by $O(\sqrt{BdFL}T^{3/4})$.*

**Proof:** We have already seen that

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

Taking expectation, using Lemma 7, and using the fact that $v_t$ is in the unit sphere, we obtain

$$E[\sum_{t=1}^{T} \hat{f}(w_t) - \hat{f}(u)] \leq \frac{1}{2\eta}||u||_2^2 + \eta \sum_{t=1}^{T} E[\frac{d^2}{\delta^2} f_t^2(w_t + \delta v_t)]$$

Then, by the Lipschitzness of $f_t$ we have $f_t(w_t+\delta v_t) \leq f_t(w_t)+L\delta||v_t|| \leq F+L\delta$. Therefore

$$E[\sum_{t=1}^{T} \hat{f}(w_t) - \hat{f}(u)] \leq \frac{B^2}{2\eta} + \eta T d^2 (F/\delta + L)^2$$

To derive a concrete bound out of the above we need to relate the regret with respect to $\hat{f}_t$ to the regret with respect to $f_t$. We do this using Lemma 6, which implies

$$f_t(w_t + \delta v_t) - f_t(u) \leq f_t(w_t) - f_t(u) + \delta L \leq \hat{f}_t(w_t) - \hat{f}_t(u) + 3\delta L$$

Combining them together we get

$$E[\sum_{t=1}^{T} f_t(w_t + \delta v_t) - f_t(u)] \leq 3L\delta T + \frac{B^2}{2\eta} + \eta T d^2 (F/\delta + L)^2$$

■

We are left with just one more issue we should address: How do we inforce that $w_t + \delta v_t$ is in $S$?
We assume that $0 \in S$ and that $U_b \subseteq S$. Since $S$ is a convex set $(1-\delta)w_t + \delta v_t \in S$, so we are left to show that the regret does not increase by too much when we look at $(1-\delta)S = \{(1-\delta)w : w \in S\}$. The following theorem will solve our issue.

**Theorem 9**

$$\min_{w \in (1-\delta)S} \sum_{t=1}^{T} f_t(w) \leq 2\delta FT + \min_{w \in S} \sum_{t=1}^{T} f_t(w)$$

**Proof:** Notice that

$$\min_{w \in (1-\delta)S} \sum_{t=1}^{T} f_t(w) = \min_{w \in S} \sum_{t=1}^{T} f_t((1-\delta)w)$$

And by the convexity of $f_t$ and the assumption that $0 \in S$ we get

$$f_t((1-\alpha)w) = f_t(\alpha 0 + (1-\alpha)w) \leq \alpha f_t(0) + (1-\alpha) f_t(w)$$

Combining them together

$$\min_{w \in (1-\delta)S} \sum_{t=1}^{T} f_t(w) \leq \min_{w \in S} \delta \sum_{t=1}^{T} f_t(0) + (1-\delta) \sum_{t=1}^{T} f_t(w) =$$

$$\min_{w \in S} \delta \sum_{t=1}^{T} f_t(0) - f_t(w) + \sum_{t=1}^{T} f_t(w) \leq 2\delta FT + \min_{w \in S} \sum_{t=1}^{T} f_t(w)$$

■

Tuning the parameters correctly will achieve the same regret bound as before.