**AI-3001: Advanced Topics in Machine Learning** **Aug-Nov 2022**

# Lecture 11: Introduction to Online Convex Optimization II

*Lecturer: Ganesh Ghalme* *Scribes: Ganesh Ghalme*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 11.1 Online Mirror Descent (OMD)

**Algorithm 1:** Online Mirror Descent (OMD)

**Input:** convex set $\mathcal{K}$, function class $\mathcal{F}$, regularization function $R$
**Initialize**: $x_1 = \arg\min_{x\in\mathcal{K}} R(x)$ and $y_1$ such that $\nabla R(y_1) = 0$ ;
**for** $t = 1, 2, \cdots$ **do**
- **Algorithm plays** $x_t \in \mathcal{K}$ ;
- **Environment reveals** $f_t \in \mathcal{F}$ ;
- **Algorithm incurs a loss** $f_t(x_t) \in \mathbb{R}$;
- **Update**

$$y_{t+1} \quad \text{such that} \quad \nabla R(y_{t+1}) = \nabla R(y_t) - \eta \nabla_t \qquad \text{[Lazy version]} \tag{11.1}$$
$$y_{t+1} \quad \text{such that} \quad \nabla R(y_{t+1}) = \nabla R(x_t) - \eta \nabla_t \qquad \text{[Agile version]} \tag{11.2}$$

- $x_{t+1} = \arg\min_{x\in\mathcal{K}} B_R(x||y_{t+1})$

**end**

### 11.1.1 Lazy version of OMD and FTRL

**Theorem 11.1.** *Let $\mathcal{F}$ be a class of linear loss functions. Then the Lazy OMD and FTRL algorithms play the same points i.e.,*

$$\arg\min_{x\in\mathcal{K}}(B_R(x||y_t)) = \arg\min_{x\in\mathcal{K}} \left(\eta \sum_{s=1}^{t-1} \nabla_s^T x + R(x)\right) \tag{11.3}$$

*Proof.* We begin the proof by observing the following

**Observation 1.** $\nabla R(y_t) = -\eta \sum_{s=1}^{t-1} \nabla_s$

The proof of above observation follows form the update rule of the Lazy version and the fact that $\nabla R(y_1) = 0$. Next, we note that

$$\nabla R(y_t) = -\eta \sum_{s=1}^{t-1} \nabla_s \tag{11.4}$$

The above equation follows from the fact that $y_t$ is unconstrained minima (first order condition). Also it is worth noting that since $R$ is strictly convex (as it is strongly convex) we have that $y_t$ is unique. [1]

Notice that in the RHS of the statement of the lemma we take projection of the point $y_t$ on $\mathcal{K}$.

$$\begin{aligned} B_R(x||y_t) = B_R(x||y_t) &= R(x) - R(y_t) - \nabla R(y_t)^T(x - y_t) \\ &= B_R(x||y_t) = R(x) - R(y_t) + \eta \sum_{s=1}^{t-1} \nabla_s (x - y_t) \end{aligned}$$

We have $\arg\min_{x\in\mathcal{K}} B_R(x||y_t) = \arg\min_{x\in\mathcal{K}}(R(x) + \eta \sum_{s=1}^{t-1} \nabla_s x)$ □

### 11.1.2 The Agile Version

We will first prove a supporting result called as generalized pythagorean inequality that holds for Bregman divergence.

**Theorem 11.2.** *Let $\mathcal{K}$ be a convex set and $x' = \Pi_{\mathcal{K}} B_R(x||y)$ be a Bregman projection of some $y \in \mathbb{R}^n$ on $\mathcal{K}$ and $u \in \mathcal{K}$. Then*

$$B_R(y||u) \geq B_R(y||x) + B_R(x||u) \tag{11.5}$$

*Proof.* First notice that from the cosine inequality for bregman divergence that it is enough to show that

$$\langle \nabla R(x') - \nabla R(x), u - x' \rangle \geq 0$$

To see why this is true, we use the following result for convex functions on a convex set.

$$\langle \nabla f(x^*), u - x^* \rangle \geq 0$$

Here $x^*$ is a minimizer of $f$ on $\mathcal{K}$ and $u \in \mathcal{K}$ . We use $f(x) = B_R(x||y)$ and note that Bregman divergence is convex in its first argument. We have

$$\begin{aligned} \nabla B_R(x^*||y) &= \nabla(R(x^*) - R(y) - \langle \nabla R(y), x^* - y \rangle) \\ &= \nabla R(x^*) - \nabla R(y) \end{aligned}$$

This completes the proof of the theorem. □

We are now ready to prove the regret guarantee of OMD.

**Theorem 11.3.** *For every $u \in \mathcal{K}$ we have*

$$\mathcal{R}_T(OMD) \leq \frac{\eta}{2} \sum_{t=1}^{T} ||\nabla||_t^{*2} + \frac{D_R^2}{\eta} \tag{11.6}$$

*Proof.* From the cosine inequality of Bregman divergence, for any $x, y, z$ we have

$$(x - y)^T(\nabla R(z) - \nabla R(y)) = B_R(x||y) + B_R(x||z) - B_R(y||z) \tag{11.7}$$

[1] Here, we point out one technicality. We define the regularization function $R$ on $\mathbb{R}^n$ and consider the regularizer to be the restriction of this function on closed and convex set $\mathcal{K} \subseteq \mathbb{R}^n$. There are multiple functions which agree over $\mathcal{K}$; take any one of them and fix that function at time $t = 1$ and all the results go through.

Further since $f_t$'s are convex (for the agile version we will consider any convex functions) we have for any $u \in \mathcal{K}$ that

$$\begin{aligned}
f_t(x_t) - f_t(u) &\leq \nabla f_t(x_t)^T(x_t - u) && \text{(Convexity of } f_t\text{'s)} \\
&= \frac{1}{\eta}(\nabla R(y_{t+1}) - \nabla R(x_t))^T(u - x_t) && \text{(choice rule of agile OMD)} \\
&= \frac{1}{\eta}(B_R(u||x_t) - B_R(u||y_{t+1}) + B_R(x_t||y_{t+1})) \\
& && \text{(Cosine inequality for Bregman Divergence)} \\
&\leq \frac{1}{\eta}(B_R(u||x_t) - B_R(u||x_{t+1}) + B_R(x_t||y_{t+1})) \\
& && (x_{t+1} \text{ is the projection of } y_{t+1} \text{ on convex set } \mathcal{K}) \\
\implies \sum_{t=1}^{T} f_t(x_t) - f_t(u) &\leq \frac{1}{\eta}\sum_{t=1}^{T}(B_R(u||x_t) - B_R(u||x_{t+1})) + \frac{1}{\eta}\sum_{t=1}^{T} B_R(x_t||y_{t+1}) \\
&= \frac{1}{\eta}\Big[B_R(u||x_1) - B_R(u||x_{T+1})\Big] + \frac{1}{\eta}\sum_{t=1}^{T} B_R(x_t||y_{t+1})
\end{aligned}$$

First, lets upper bound the first term in the RHS. We have

$$\begin{aligned}
B_R(u||x_1) - B_R(u||x_{T+1}) &\leq B_R(u||x_1) && \text{(since } B_R(.||.) \text{ is non-negative)} \\
&= R(u) - R(x_1) - \nabla R(x_1)^T(u - x_1) \\
&\leq R(u) - R(x_1) && (\ x_1 = \arg\min_{x \in \mathcal{K}} R(x)) \\
&\leq D_R^2
\end{aligned}$$

Next, we bound the second term as follows.

$$\begin{aligned}
B_R(x_t||y_{t+1}) + B_R(y_{t+1}|x_t|) &= R(x_t) - R(y_{t+1}) - \nabla R(y_{t+1})^T(x_t - y_{t+1}) + R(y_{t+1}) - R(x_t) - \nabla R(x_t)^T(y_{t+1} - x_t) \\
&= (\nabla R(x_t) - \nabla R(y_{t+1}))^T(x_t - y_{t+1}) \\
&\leq \eta \sum_{t=1}^{T} \nabla f_t(x_t)^T(x_t - y_{t+1}) && \text{(update rule of Agile version)} \\
&\leq \eta ||\nabla f_t(x_t)||_t^* ||x_t - y_{t+1}||_t && \text{(generalized Cauchy-Shwartz inequality)} \\
&\leq \frac{\eta^2 ||\nabla_t||_t^{\star 2}}{2} + \frac{||x_t - y_{t+1}||_t^2}{2} && \text{(AM-GM inequality)}
\end{aligned}$$

This implies

$$\begin{aligned}
B_R(x_t||y_{t+1}) &\leq \frac{\eta^2 ||\nabla_t||_t^{\star 2}}{2} + \frac{||x_t - y_{t+1}||_t^2}{2} - B_R(y_{t+1})||x_t) \\
&= \frac{\eta^2 ||\nabla_t||_t^{\star 2}}{2} \\
& \text{(from the definition of Bregman Divergence } B_R(y_{t+1})||x_t) = \tfrac{||x_t - y_{t+1}||_t^2}{2})
\end{aligned}$$

Replacing above inequalities in the upper bound of regret completes the proof. □

The Bregman divergence preserves all the properties (that is convexity and pythagorean property) required for the projection step to be well defined. We next give a few remarks about OMD and Mirror Descent in-general.

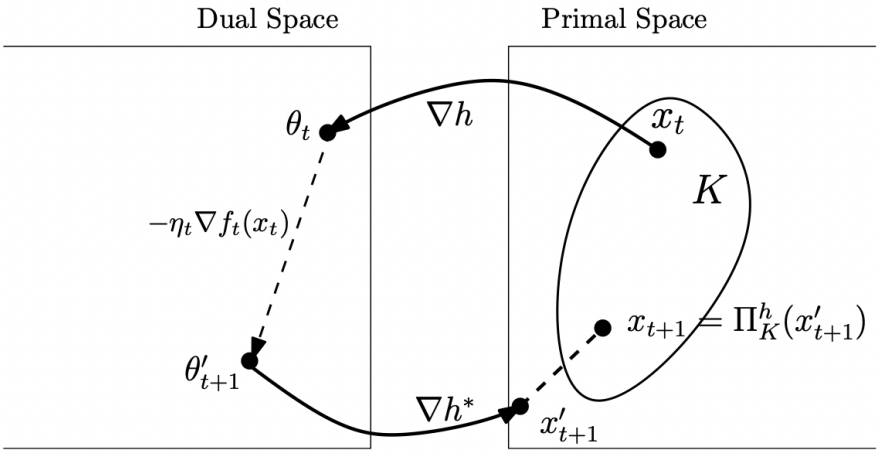


Figure 11.1: Figure taken for reference from [1]

## 11.2 Remarks

- The dual role of regularization function; as a stability parameter and as a *teleportation device*: We started with FTL algorithm that, in a given example, picked *extreme* points whereas the optimal strategy was to pick a midpoint. This lead to linear regret. We first introduced the regularizer in FTRL to fix this issue. However, in OMD we use strongly convex regularizer to *mirror* the points in original $x$-space to a gradient space and back. We saw that when we return to the $x$-space we may not land in a convex set. Bregman projection provided us a tool to fix this issue.

- OMD significantly generalizes the projected (sub)-gradient descent where we find

  $x_{t+1} = \arg\min_{x \in \mathcal{K}} 1/2||x - (x_t - \eta \nabla_t)||_2^2$. Clearly, we recover the projected gradient descent by considering $R(x) = 1/2||x||_2^2$; however, by doing so, we are stuck in euclidean geometry and the bounds that depend on radius and girth (upper bound on (dual) norms of gradients) could be loose. With regularizers, we can do much better. We now get to incorporate some side information (say we know the bound on infinity norm of the gradient, or the directional derivative of functions or nature of the convex set) while choosing our regularization function. In this sense, OMD strictly generalizes gradient descent.

- Mirror Descent algorithm in more general and finds applications beyond regret minimization framework and beyond online setting.

- This technique is surprisingly robust. If we gets corrupted data (for instance gradient directions are random but matches true gradient in expectation), the expected performance does not deteriorate rapidly. We can use Stochastic MIrror descent and work with the (reasonable) estimates and we are good. Even if someone corrupts a few ($\varepsilon$ fraction) of gradients completely, the loss in the performance is affected just by $\varepsilon$.

- It works with less information. That is it only needs to know the grandients and not the entire function to make its updates. This is better than the FTRL (the original version).

- Perhaps the only drawback is the projection step. We must project $B_R(.||.)$ on the convex set and this projection could be computationally costly for some $R$ (maybe for optimal $R$). One needs to be careful.

- There are multiple proofs of OMD in literature. I loosely followed the proof given in OCO book (Chapter 5). For alternate proof that is based on the potential function argument is given in [1]. Sebastian Bubeck has a great video lectures title *Five miracles of Mirror Descent* (`https://www.youtube.com/watch?v=5DIZCxcfeWU`)

## References

[1] Potential-Function Proofs for First-Order Methods, Nikhil Bansal and Anupam Gupta, arXiv, 2017, `https://arxiv.org/abs/1712.04581`,