IE613: Online Machine Learning **Jan-Apr 2018**

Lecture 16: Mirror Descent

Lecturer: M. K. Hanawal *Scribes: Shubham Uttam*

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.*

16.1 Recap

In the last class we saw the Online gradient descent algorithm

Algorithm 1 Online Gradient Descent

1: Start with $w_1 \in \mathcal{K} \subset \mathbb{R}^d$
2: **for** `rounds t-1,....,T` **do**
3: `player plays` w_t
4: `environment plays` w_t
5: $w_{t+1} \leftarrow Proj_k[w_t - \eta_t \bigtriangledown c_t(w_t)]$

Note that here the set $\mathcal{K}$ is closed and bounded and the cost function c_t is strongly convex and continuously differentiable. When c_t is convex but not continuously differentiable then we get online sub-gradient descent algorithm by making some changes in the update statement.

Algorithm 2 Online Sub-Gradient Descent

1: Start with $w_1 \in \mathcal{K} \subset \mathbb{R}^d$
2: **for** `rounds t-1,....,T` **do**
3: `player plays` w_t
4: `environment plays` w_t
5: $w_{t+1} \leftarrow Proj_k[w_t - \eta_t \quad \bigtriangledown c_t(w_t)]$

16.2 Some basics of Linear Algebra

Definition 16.1 ***Field:*** *A field F is a set with two operations addition and multiplication,*

$$+ : F \times F \rightarrow F \text{ and } . : F \times F \rightarrow F$$

which obey the following axioms:

- *(F,+) is an* ***abelian group*** *under addition:*

(1) Addition is associative. That is for every x, y and z $\in F$,

$$(x + y) + z = x + (y + z).$$

(2) *There is an identity element under addition. This element is often denoted 0$\in F$ and for every element x$\in F$,*

$$0 + x = x + 0 = x.$$

(3) *Every element has an additive inverse. That is given x$\in F$, there is an element -x$\in F$ and*

$$x + (-x) = x - x = 0$$

(4) *Addition is commutative. That is given x and y $\in F$,*

$$x + y = y + x.$$

- *Let F*=F-{0}. Then (F*,.) is an abelian group under multiplication:*

(1) *Multiplication is associative. That is for every x, y and z$\in F$,*

$$(x.y).z = x.(y.z)$$

(2) *There is an identity element under multiplication. This element is often denoted 1$\in F$ and for every element x$\in F$,*

$$1.x = x.1 = x$$

(3) *Every element has an multiplicative inverse. That is given x$\in F$, there is an element $x^{-1} \in F$ and*

$$x.x^{-1} = x^{-1}.x = 0$$

(4) *multiplication is commutative. That is given x and y $\in F$,*

$$x.y = y.x$$

- *addition and multiplication are compatible, i.e. F satisfies the distributive law. that is given x, y and z$\in F$,*

$$x.(y + z) = x.y + x.z$$

Definition 16.2 ***Vector Space:*** *A vector space over a field F is a set V together with two operations, vector addition and vector multiplication that satisfy the eight axioms:*
Let u,v,w $\in V$ and a,b$\in F$

- *Associativity of addition:*

$$u + (v + w) = (u + v) + w$$

- *Commutativity of addition:*

$$u + v = v + u$$

- *Identity element of addition:*
  there exists an element 0$\in V$, called zero vector, such that v+0 =v for all v $\in V$

- *Inverse elements of addition:*
  For every v $\in$ V, there exists an element v $\in$ V, called the additive inverse of v, such that v + (-v) = 0.

- *Compatibility of scalar multiplication with field multiplication:*

$$a(bv) = (ab)v$$

- *Identity element of scalar multiplication:*
  1v=v, where 1 denotes the multiplicative identity in F.
- *Distributivity of scalar multiplication with respect to vector addition:*

$$a(u + v) = au + av$$

- *Distributivity of scalar multiplication with respect to field addition:*

$$(a + b)v = av + bv$$

Definition 16.3 ***Inner Product Space:*** *An inner product space is a vector space V over the field F together with an inner product, i.e., with a map*

$$< \cdot, \cdot >: V \times V \to F$$

that satisfies the following three axioms for all vectors x,y,z∈ V and all scalars a∈ F:

- *Conjugate symmetry:*

$$< x, y >= \overline{< y, x >}$$

- *Linearity in the first argument:*

$$< ax, y >= a < x, y >$$

$$< x + y, z >=< x, z > + < y, z >$$

- *Positive-definiteness:*

$$< x, x >\geq 0$$

$$< x, x >= 0 \Leftrightarrow x = 0$$

Definition 16.4 ***Complete Space:*** *A metric space M is called complete (or a Cauchy space) if every Cauchy sequence of points in M has a limit that is also in M or, alternatively, if every Cauchy sequence in M converges in M.*

Definition 16.5 ***Hilbert Space:****A complete space with an inner product is called a Hilbert space.*

Definition 16.6 ***Linear Functional:*** *A map $f : V \to F$ over a vector space V over F is called a linear functional if $f(\alpha u + \beta v) = \alpha f(u) + \beta f(v)$, $\forall u, v \in V$, $\forall \alpha, \beta \in F$.*

Definition 16.7 ***Dual Space:*** *Let H be a Hilbert space. Then the set of all linear functional over H forms the dual space H*.*

Riesz Representation Theorem:
Let H be a Hilbert space, and let H^* be its dual space. If x is an element of H, then the function φ_x, for all y in H defined by:

$$\varphi_x(y) =< y, x >$$

Theorem: The mapping $\phi : H \to H^*$ defined by $\phi(x) = \varphi_x$ is an isometric isomorphism, meaning that:

- ϕ is bijective.
- The norm of x and φ_x agree: $||x|| = ||\phi(x)||$.
- ϕ is additive: $\phi(x_1 + x_2) = \phi(x_1) + \phi(x_2)$.
- If the base field is $\mathbb{R}$, then $\phi(\lambda x) = \lambda\phi(x)$ for all real numbers λ.
- If the base field is $\mathbb{C}$, then $\phi(\lambda x) = \bar{\lambda}\phi(x)$ for all complex numbers λ, where $\bar{\lambda}$ denotes the complex conjugation of λ.

Definition 16.8 ***Banach space:*** *A Banach space is a vector space X over the field $\mathbb{R}$, or over the field $\mathbb{C}$, which is equipped with a norm and which is complete with respect to that norm, that is to say, for every Cauchy sequence $\{x_n\}$ in X, there exists an element x in X such that*

$$\lim_{n\to\infty} x_n = x,$$

or equivalently:

$$\lim_{n\to\infty} ||x_n - x||_x = 0$$

If B is Banach space, then Riesz Representation theorem may not hold.
If g is a linear functional on $\mathbb{R}^d$, then:

$$g(y) = y_1 g(e_1) + y_2 g(e_2) + ... + y_d g(e_d)$$

where $y = y_1e_1 + y_2e_2 + ... + y_de_d$ where $\{e_1, e_2, ..., e_d\}$ is the set of basis of $\mathbb{R}^d$.

16.3 Mirror Descent

Definition 16.9 ***Bregman Divergence:*** *Let $f : \omega \to \mathbb{R}$ be a function that is strictly convex, continuously differentiable and defined on a closed convex set ω. Then the Bregman divergence is defined as:*

$$B_f(x, y) := f(x) - f(y) - < \bigtriangledown f(y), x - y >,\ \forall x, y \in \omega$$

That is, the difference between the value of f at x and the first order Taylor expansion of f around y evaluated at point x.

Examples of Bregman divergence

- Euclidean distance. Let $f(x) = \frac{1}{2}||x||^2$. Then Bregman divergence $B_f(x, y) = \frac{1}{2}||x - y||^2$
- Non Euclidean distance, Mahalanobis distance. We have $f(x) = \frac{1}{2}x^TAx$, for some positive semidefinite matrix A. Then Bregman divergence is

$$B_f(x, y) = \frac{1}{2}(x - y)^T A(x - y)$$

- Kullback-Leibler divergence. We have $f(p) = \sum p_i \log p_i$. Then Bregman divergence is

$$B_f(p, q) = \sum p_i \log \frac{p_i}{q_i} - \sum p_i + \sum q_i$$

Two components which we will use in mirror descent procedure:

1. Mirror function, ϕ
   It is a convex function defined as, $\phi : V \to V^*$ where V^* is the dual space of V.

2. Bregman projection

$$Proj_K(u) = \arg\min_{v \in K} B_f(v, u)$$

Definition 16.10 ***Fenchel-Legendre dual:*** *A funtion $f : \mathcal{H} \to \mathbb{R}$, define $f^* : \mathcal{H}^* \to \mathbb{R}$ such that*

$$f^*(u) = \sup_{x \in \mathcal{H}} \{f(x) - < u, x >\}$$

If function f is convex and differentiable, then $f^*(\bigtriangledown f(x)) = f(x) - < \bigtriangledown f(x), x >$

Motivation of mirror descent procedure: If there is a case where $\bigtriangledown c_t(w_t) \notin K$, ($K \neq \mathbb{R}^d$) then nothing can be said about $w_t - \eta_t \bigtriangledown c_t(w_t)$, so we conduct update in dual space instead of real space.

16.3.1 Online Mirror Descent Procedure

$\to$ Update in the dual space
$\to$ Take a mirror function, ϕ.
$\to$ Mirror Update: $\bigtriangledown \phi(\widetilde{w}_{t+1}) \leftarrow \bigtriangledown \phi(w_t) - \eta_t \bigtriangledown c_t(w_t)$
$\to$ $w_{t+1} \leftarrow$ Bregman$Proj_K^{\phi}[\widetilde{w}_t]$

If we take $\phi(x) = \frac{1}{2}||x||^2$, then it becomes Online Descent Algorithm and if we take ϕ to be a negative entropy term in mirror update then algorithm we get is known as Online Exponential Gradient algorithm.

16.3.2 Regret of OMD

Let environment reveals a convex regret loss function as c_t in round t, then the regret for online mirror descent is given by:

$$Regret_{OMD}(u, T) = \sum_{t=1}^{T} c_t(w_t) - \sum_{t=1}^{T} c_t(u)$$

16.3.3 Regret Bound for OMD

$$
\begin{aligned}
\text{Consider } c_t(w_t) - c_t(u) &\leq < \triangledown c_t(w_t), w_t - u > \quad \text{(By convexcity of } c_t\text{)} \\
&= \frac{1}{n} < \triangledown \phi(w_t) - \triangledown \phi(\widetilde{w}_{t+1}), w_t - u > \quad (\text{ since } \triangledown \phi(\widetilde{w}_{t+1}) = \triangledown \phi(w_t) - \eta \triangledown \phi(w_t)) \\
&= \frac{1}{n}[B_\phi(u, w_t) - B_\phi(u, \widetilde{w}_{t+1}) + B_\phi(w_t, \widetilde{w}_{t+1})] \quad (\text{ using Bregman divergence}) \\
&= \frac{1}{n}[B_\phi(u, w_t) - B_\phi(w_{t+1}, \widetilde{w}_{t+1}) - B_\phi(u, w_{t+1}) + B_\phi(w_t, \widetilde{w}_{t+1})] \\
&\quad (\text{as } B_\phi(u, \widetilde{w}_{t+1}) = B_\phi(w_{t+1}, \widetilde{w}_{t+1}) + B_\phi(u, w_{t+1}))
\end{aligned}
$$

Summing over T:

$$
\begin{aligned}
\sum_{t=1}^{T}(c_t(w_t) - c_t(u)) &\leq \frac{1}{n}\sum_{t=1}^{T}[B_\phi(u, w_t) - B_\phi(u, w_{t+1})] + \frac{1}{n}\sum_{t=1}^{T}[B_\phi(w_t, \widetilde{w}_{t+1}) - B_\phi(w_{t+1}, \widetilde{w}_{t+1})] \\
&= \frac{1}{n}[B_\phi(u, w_1) - B_\phi(u, w_{T+1})] + \frac{1}{n}\sum_{t=1}^{T}[B_\phi(w_t, \widetilde{w}_{t+1}) - B_\phi(w_{t+1}, \widetilde{w}_{t+1})] \\
&\leq \frac{1}{n}[B_\phi(u, w_1)] + \frac{1}{n}\sum_{t=1}^{T}[B_\phi(w_t, \widetilde{w}_{t+1}) - B_\phi(w_{t+1}, \widetilde{w}_{t+1})] \\
&\quad (\text{as for some convex } \phi, B_\phi(u, v) \geq 0)
\end{aligned}
$$

For $w_1 : \phi(u) - \phi w_1 \leq D^2, \quad \forall u \in K$

$B_\phi(u, w) = \phi(u) - \phi(w_1) - < \triangledown \phi(w_1), u - w_1 > \leq D^2 \quad$ (as $< \triangledown \phi(w_1), u - w_1 > \geq 0$)

Therefore we get:

$$
\begin{aligned}
\sum_{t=1}^{T}(c_t(w_t) - c_t(u)) &\leq \frac{1}{n}D^2 + \frac{1}{n}\sum_{t=1}^{T}[B_\phi(w_t, \widetilde{w}_{t+1}) - B_\phi(w_{t+1}, \widetilde{w}_{t+1})] \\
&= \frac{1}{n}D^2 + \frac{1}{n}\sum_{t=1}^{T}[\phi(w_t) - \phi(w_{t+1}) - < \triangledown \phi(\widetilde{w}_{t+1}), w_t - w_{t+1} >]
\end{aligned}
$$

Assumption: ϕ is strongly convex with modulus α

$i.e.\ \phi(y) \geq \phi(x) + < \triangledown \phi(x), y - x > + \frac{\alpha}{2}||y - x||^2$

$$
\text{Therefore, } \sum_{t=1}^{T}(c_t(w_t) - c_t(u)) \leq \frac{1}{\eta}D^2 + \sum_{t=1}^{T}[< \triangledown \phi(w_t), w_t - w_{t+1} > - \frac{\alpha}{2}||w_t - w_{t+1}||^2 - < \triangledown \phi(\widetilde{w}_{t+1}), w_t - w_{t+1} >]
$$

Recall update step:

$$
\begin{aligned}
\triangledown \phi(\widetilde{w}_{t+1}) &= \triangledown \phi(w_t) - \eta \triangledown c_t(w_t) \\
\triangledown \phi(w_t) - \triangledown \phi(\widetilde{w}_{t+1}) &= \eta \triangledown c_t(w_t)
\end{aligned}
$$

$$
\begin{aligned}
\text{Therefore, } \sum_{t=1}^{T}(c_t(w_t) - c_t(u)) &\leq \frac{1}{\eta}D^2 + \sum_{t=1}^{T}[< \triangledown \phi(w_t) - \triangledown \phi(\widetilde{w}_{t+1}), w_t - w_{t+1} > - \frac{\alpha}{2}||w_t - w_{t+1}||^2] \\
&= \frac{1}{\eta}D^2 + \sum_{t=1}^{T}[\eta < \triangledown c_t(w_t), w_t - w_{t+1} > - \frac{\alpha}{2}||w_t - w_{t+1}||^2] \\
&= \frac{1}{\eta}D^2 + \sum_{t=1}^{T}[< \eta \triangledown c_t(w_t), w_t - w_{t+1} > - \frac{\alpha}{2}||w_t - w_{t+1}||^2]
\end{aligned} \tag{16.1}
$$

Now by Cauchy Schwartz Inequality, we have:

$$| < x, y > |^2 \leq ||x|| \; ||y||$$

So then we have:

$$< \eta \bigtriangledown c_t(w_t), w_t - w_{t+1} > = \eta < \bigtriangledown c_t(w_t), w_t - w_{t+1} >$$
$$< \eta \bigtriangledown c_t(w_t), w_t - w_{t+1} > \leq \eta || \bigtriangledown c_t(w_t)|| \; ||w_t - w_{t+1}||$$

Now as $|| \bigtriangledown c_t(w_t)||$ is bounded by G, we have

$$< \eta \bigtriangledown c_t(w_t), w_t - w_{t+1} > \leq \eta G \; ||w_t - w_{t+1}||$$

So we have:

$$< \eta \bigtriangledown c_t(w_t), w_t - w_{t+1} > - \frac{\alpha}{2}||w_t - w_{t+1}||^2 \leq \frac{\eta^2 G^2}{2\alpha} + \frac{\alpha}{2}||w_t - w_{t+1}||^2 - \eta G||w_t - w_{t+1}||^2$$
$$= (\frac{\eta G}{\sqrt{2\alpha}} - \sqrt{\frac{\alpha}{2}}||w_t - w_{t+1}||)^2$$

Therefore the regret bound we have is:

$$Reg = \sum_{t=1}^{T}(c_t(w_t) - c_t(u)) \leq \frac{1}{\eta}D^2 + \sum_{t=1}^{T}\frac{\eta^2 G^2}{2\alpha} = \frac{1}{\eta}[D^2 + \frac{\eta^2 G^2 T}{2\alpha}]$$

16.4 Follow the Regularized Leader

Follow the regularized leader minimizes the loss of all past rounds plus a regularization term. The goal of the regularization term is to stabilize the solution. Formally, for a regularization function, $R : S \rightarrow \mathbb{R}$ we define the weight for any round t as:

$$w_t = arg \min_{w \in S} \sum_{i=1}^{t-1} f_i(w) + R(w)$$