Statistical Techniques in Robotics (16-831, S22) Lecture #08 (Monday, February 14)

Online Mirror Descent

Lecturer: Kris Kitani *Scribes: Feng Xiang, Yuqing Qin*

1 Review

In the previous lecture, we introduced the "Follow the Leader" algorithm with quadratic loss and "Follow the Regularized Leader" with linear loss and quadratic regularization. We also proved both algorithms are no-regret algorithms.

1.1 FTL with Quadratic Loss

The quadratic loss is defined as below:

$$f^{(t)}(w) = \frac{1}{2}||\boldsymbol{w} - \boldsymbol{z}^{(t)}||_2^2$$

The algorithm is summarized as below:

Algorithm 1 Follow the Leader with Quadratic Loss

```
for $t = 1, 2, \cdots, T$ do
    $\boldsymbol{w}^{(t)} = \underset{\boldsymbol{w} \in W}{\arg\min} \sum_{i=1}^{t-1} f^{(i)}(\boldsymbol{w})$        ▷ Parameter estimation
    RECEIVE $(f^{(t)}(w) = \frac{1}{2}||\boldsymbol{w} - \boldsymbol{z}^{(t)}||_2^2)$        ▷ Quadratic loss function
end for
```

With the quadratic loss, and the proper assumption we made, we have seen that the regret bound is:

$$Regret \leq 4L^2(log(T) + 1)$$

From the regret bound, we could see that the FTL with quadratic loss is a no-regret algorithm since the regret bound is sub-linear with a big-O time complexity of $O(logT)$.

1.2 FTRL with Linear Loss and Quadratic Regularization

The general version of the Follow the Regularized Leader is just like FTL with a regularizer term added in the parameter estimation. We can even think of FTL as a special case of FTRL where $\psi = 0$. To note, S is a convex set shown in Algorithm 2 below.

Algorithm 2 Follow the Regularized Leader

```
for $t = 1, 2, \cdots, T$ do
    $\boldsymbol{w}^{(t)} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{t-1} f^{(i)}(\boldsymbol{w}) + \psi(\boldsymbol{w})$        ▷ Parameter estimation
    RECEIVE $(f^{(t)} : S \to \mathbb{R})$        ▷ Update loss function
end for
```

The quadratic regularization term is defined as:

$$\psi(\boldsymbol{w}) = \frac{1}{2\eta}||\boldsymbol{w}||_2^2$$

To calculate a linear loss, the following function is used:

$$f^{(t)} = \boldsymbol{w} \cdot \boldsymbol{z}^{(t)}$$

With the regret bound being proved as:

$$R^{(T)}(\boldsymbol{u}) \leq BL\sqrt{2T}$$

Where the term definitions are:

$$L = max_{\boldsymbol{z}}||\boldsymbol{z}||_2$$

$$B = max_{\boldsymbol{u} \in S}||\boldsymbol{u}||_2$$

We could see the FTRL with linear loss and quadratic regularization is also a no-regret algorithm.

In today's lecture, we will introduce another Online Convex Optimization algorithm, Online Mirror Descent (OMD). The FTRL can be interpreted as the OMD algorithm.

2 Summary

2.1 Online Mirror Descent (OMD)

As seen in previous lectures, FTRL is a generic algorithm for convex optimization. OMD is another useful framework that could connect many online learning algorithms. This alternative perspective gives us a unification of online learning algorithms and more mathematical tools to analyze regret bound. In today's lecture, we will start by deriving the OMD algorithm from the FTRL-LinLoss algorithm we learned last time, and then derive the general regret bound for the OMD algorithm.

2.2 Derive OMD from FTRL-LinLoss

FTRL with linear loss function and convex regularization algorithm is summarized in Algorithn 3 shown below.

Algorithm 3 FTRL-LinLoss

```
for $t = 1, 2, \cdots, T$ do
    $\boldsymbol{w}^{(t)} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{t-1} \boldsymbol{w} \cdot \boldsymbol{z}^{(i)} + \psi(\boldsymbol{w})$        ▷ Linear loss and regularization
    Receive $(f^{(t)} : S \to \mathbb{R})$        ▷ Update loss function
end for
```

We first start by generalizing the FTRL linear loss parameter sum. The one we are familiar with is:

$$z^{(1:t)} = \sum_{i=1}^{t} z^{(i)}$$

The above equation is the sum of gradients of linear loss, we could also define the sum of gradients in the dual space. The dual space parameter θ to define the following:

$$\theta^{(t+1)} = -z^{(1:t)}$$

The above definition could be written into an incremental sum:

$$\theta^{(t+1)} = \theta^{(t)} - z^{(t)}$$

The second step is to generalize the FTRL prediction. As shown in the algorithm above, the weight is defined as:

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \arg\min_{\boldsymbol{w}} \sum_{i=1}^{t} f^{(i)}(w) + \psi(\boldsymbol{w}) \\ &= \arg\min_{\boldsymbol{w}} \sum_{i=1}^{t} \langle w, z^{(1:t)} \rangle + \psi(\boldsymbol{w}) \end{aligned}$$

Multiply -1, and revert the min to max, the follow equation summarized to:

$$= \arg\max_{\boldsymbol{w}} \sum_{i=1}^{t} \langle w, -z^{(1:t)} \rangle - \psi(\boldsymbol{w})$$

We call the w to be the parameter of the primal space. We could further substitute the dual space parameter into the equation:

$$= \arg\max_{\boldsymbol{w}} \sum_{i=1}^{t} \langle w, \theta^{(t+1)} \rangle - \psi(\boldsymbol{w})$$

This equation can be summarized as a linking function that map the dual parameter (θ) to the primal parameter (w):

$$\boldsymbol{w}^{(t+1)} = g(\theta^{(t+1)})$$

We call this mapping function g to be the 'mirror/linking function', which maps the dual space to primal space. g is connecting the dual space and primal space:

$$\boldsymbol{w}^{(t+1)} = g(\theta^{(t+1)}) = \arg\max_{\boldsymbol{w}} \sum_{i=1}^{t} \langle w, \theta^{(t+1)} \rangle - \psi(\boldsymbol{w})$$

From the above derivation, we could generalize the FTRL-LinLoss to be OMD. The OMD algorithm is shown to be the following:

Algorithm 4 Online Mirror Descent

```
for $t = 1, 2, \cdots, T$ do
    RECEIVE $(f^{(t)} : \mathcal{S} \to \mathbb{R})$
    $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \boldsymbol{z}^{(t)}$, $\boldsymbol{z} \in \partial f^{(t)}(\boldsymbol{w}^{(t)})$        ▷ Dual parameter update
    $\boldsymbol{w}^{(t+1)} = g(\boldsymbol{\theta}^{(t+1)})$        ▷ Primal parameter update
end for
```

OMD is a generic algorithm for solving the online convex optimization problem. Specifically, the above OMD algorithm focuses on linear loss and convex regularizer cases.

To better understand the OMD, the geometry of the functions might help. The Figure 1 is a simulation of the online mirror descent process. The online gradient descent (which we are familiar with) is directly updating the primal parameter (w) in primal space by taking one gradient step. However, for online mirror descent, instead of updating the parameter in primal space, we could optimize the parameter in dual space (θ). Then we could project it to the primal space and get the primal parameter (w).

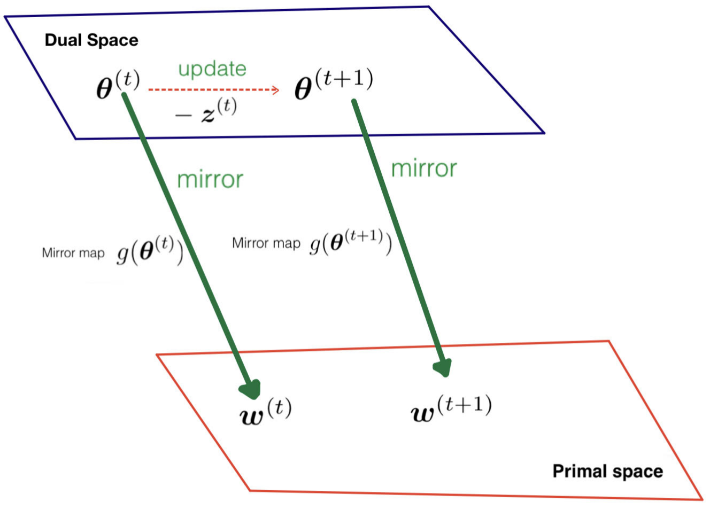
Dual Space
$\boldsymbol{\theta}^{(t)}$
update
$-\boldsymbol{z}^{(t)}$
$\boldsymbol{\theta}^{(t+1)}$
mirror
mirror
Mirror map $g(\boldsymbol{\theta}^{(t)})$
Mirror map $g(\boldsymbol{\theta}^{(t+1)})$
$\boldsymbol{w}^{(t)}$
$\boldsymbol{w}^{(t+1)}$
Primal space

Figure 1: Illustration of online mirror descent process

One thing to keep in mind is that different regularizations will lead to different mirror functions. We ought to choose good g functions to take better advantage of the geometry of the solution space.

2.3 Duality

The overall topic of duality within this context relates to the phrase: ”If the dual of A is B, then the dual of B is A.” The subtopics of duality are convex conjugates and Bregman Divergence, two mathematical tools used to understand the OMD regret bound calculation.

2.3.1 Convex Conjugate

Given a primal set (i.e. parameter, function set) $\{\psi(w), w\}$, there exists a dual set expressed in slope inputs and intercept functions $\{b(\theta), \theta\}$. The traditional method of characterizing functions is in the primal set, where the input w is mapped to a function $\psi(w)$. In the same sense of characterizing functions, one can also describe a function based on a set of line equations at each point such that each slope θ has a mapping function to output intercepts $b(\theta)$. A visualization of the dual set is shown in Figure 2 below. Given the dual equation at point w^* is $\psi(w^*) =< \theta, w^* > -b$, then one can determine the intercept as a function of θ at w^* such that $-b(\theta) = - < \theta, w^*(\theta) > +\psi(w^*(\theta))$.

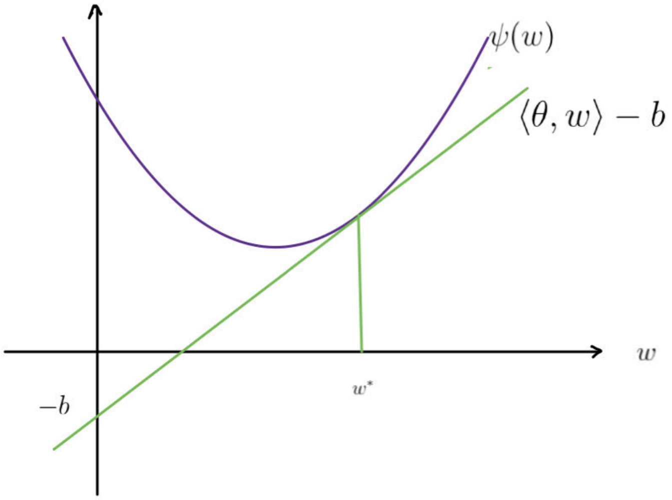

Figure 2: Dual set visualization

Now suppose one draws the line $< \theta, w > -\psi(w)$ through the origin, one can visualize the vertical incremental distance between points of $\psi(w)$ and the line. Computing the argument that would output the maximum distance between the sloped line and $\psi(w)$ turns out to be $w*$, or the input argument for the sloped line. A visualization is shown in Figure 3 below. The value of that maximum distance also turns out to be $\psi^*(\theta) = \max_w(< \theta, w > -\psi(w))$. Thus, the definition of the convex conjugate function is: $\psi^*(\theta) = \max_w(< \theta, w > -\psi(w))$.

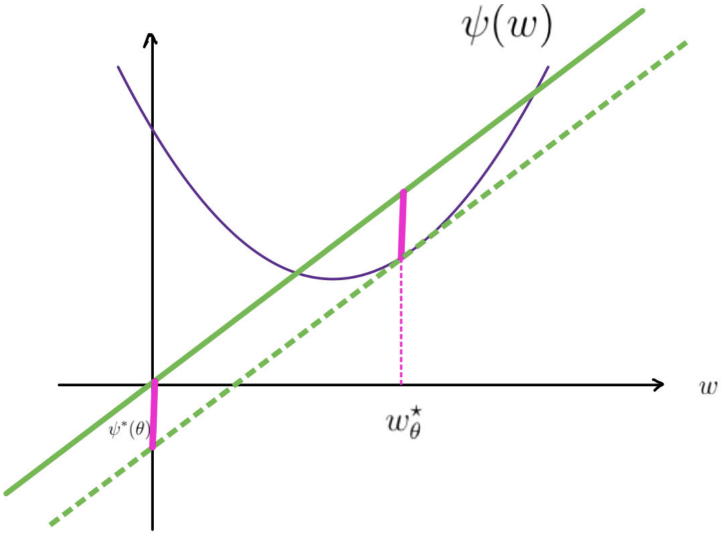

Figure 3: Dual set argmax illustration

The usefulness of the convex conjugate function $\psi^*(\theta) = \max_w(< \theta, w > -\psi(w))$ comes with the various mathematical properties inherent in such a form [2]:

1. The derivative of the complex conjugate with respect to θ is equal to the optimal point of the

function $\psi(w^*)$: $\nabla_\theta \psi^*(\theta) = \frac{\partial \psi^*(\theta)}{\partial \theta} = w^*$

2. The derivative of the convex function with respect to w is equal to the slope θ: $\nabla_w \psi(w) = \frac{\partial \psi(w)}{\partial w}|_{w=w^*} = \theta$

3. The complex conjugate computed for a specific θ input is lower bounded by the line equation with any other input w (Fenchel-Young inequality): $\psi^*(\theta) \geq (<\theta, w> - \psi(w))$

2.3.2 Bregman Divergence

From a high-level overview, the Bregman Divergence describes the approximation error between a linear approximation and a convex function between two input points u and w.

Let Figure 4 shown below be a visualization for the Bregman Divergence. Given two points along a convex function, the difference between the two values can be seen as the incremental rise along the line function from point w to point u in addition to some divergence term $D_\psi(w||u)$. Thus, the divergence term (i.e. Bregman Divergence) can be derived to be [1]:

$$\psi(w) - \psi(u) = \nabla\psi(u)^T(w-u) + D_\psi(w||u)$$
$$D_\psi(w||u) = \psi(w) - \psi(u) - \nabla\psi(u)^T(w-u)$$

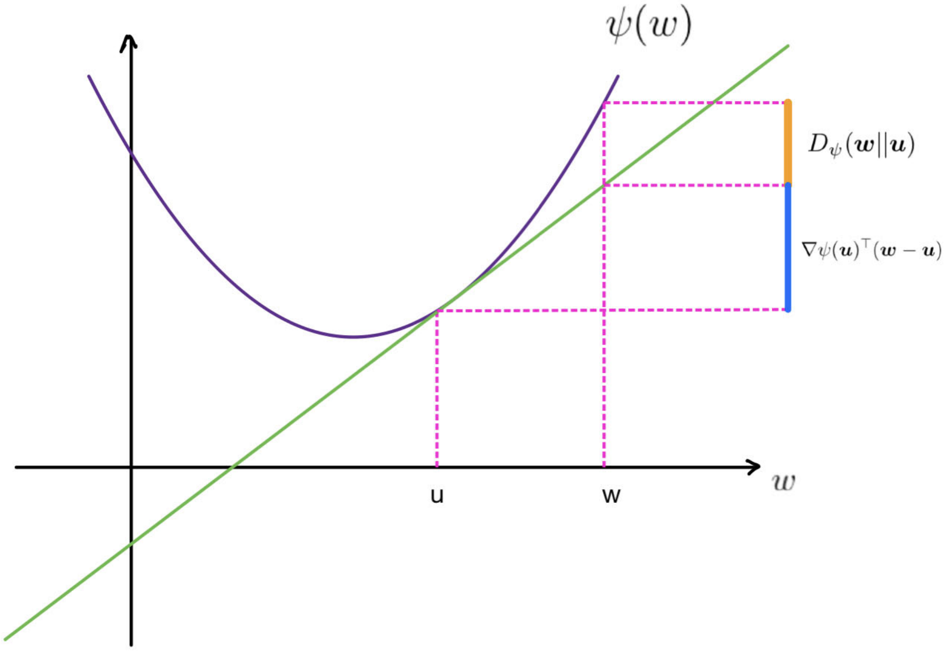
$\psi(w)$
$D_\psi(\boldsymbol{w}||\boldsymbol{u})$
$\nabla\psi(\boldsymbol{u})^\top(\boldsymbol{w}-\boldsymbol{u})$
u
w
w

Figure 4: Bregman divergence graph illustration

2.4 OMD Regret Bound Analysis

Given the aforementioned concepts of convex conjugates and the Bregman Divergence, we can now derive the regret bound for the OMD algorithm. The starting equation is the generic regret algorithm given by the following:

$$R(u) = \sum_{t=1}^{T} w^{(t)} \cdot z^{(t)} - u \cdot z^{(t)}$$
$$R(u) = \psi(u) + \sum_{t=1}^{T} u \cdot z^{(t)}$$
$$R(u) = \psi(u) - u \cdot \theta^{(T+1)}$$

Applying the Fenchel-Young inequality gives us the following lower bound:

$$\psi(u) - u \cdot \theta^{(T+1)} \geq -\psi^*(\theta^{(T+1)})$$

Then, apply telescoping to the right-hand side of the inequality introduces the Bregman Divergence term. In the last line of the derivation below, the right hand side is broken up into two terms, where the second term is considered the first-order approximation of $\psi(w)$ and the third term is the Bregman Divergence.

$$-\psi^*(\theta^{(T+1)}) = -\psi^*(\theta^{(T+1)}) - \psi^*(\theta^{(1)}) + \psi^*(\theta^{(1)})$$
$$-\psi^*(\theta^{(T+1)}) = -\psi^*(\theta^{(T+1)}) - \psi^*(\theta^{(T)}) + \psi^*(\theta^{(T)}) - \cdots - \psi^*(\theta^{(1)}) + \psi^*(\theta^{(1)})$$
$$-\psi^*(\theta^{(T+1)}) = -\psi^*(\theta^{(1)}) - \sum_{t=1}^{T} (\psi^*(\theta^{(t)}) - \psi^*(\theta^{(t)}))$$
$$-\psi^*(\theta^{(T+1)}) = -\psi^*(\theta^{(1)}) - \underbrace{\sum_{t=1}^{T} (\nabla\psi^*(\theta^{(t)}) \cdot (\theta^{(t+1)} - \theta^{(t)})}_{1^{st} \text{ order approx. of } \psi(w)} + \underbrace{D_{\psi^*}(\theta^{(t+1)}||\theta^{(t)})}_{\text{Bregman Divergence}}$$

Substituting the definition of the convex conjugate to be $\psi^*(\theta^{(1)}) = -\psi(w^{(1)})$ and the convex conjugate derivative property $\nabla_\theta \psi^*(\theta) = w^*$, the equation becomes:

$$-\psi^*(\theta^{(T+1)}) = -\psi^*(\theta^{(1)}) - \sum_{t=1}^{T}(\nabla\psi^*(\theta^{(t)}) \cdot (\theta^{(t+1)} - \theta^{(t)}) + D_{\psi^*}(\theta^{(t+1)}||\theta^{(t)})$$

$$-\psi^*(\theta^{(T+1)}) = -\psi(w^{(1)}) - \sum_{t=1}^{T}(w^{(t)} \cdot (-z^{(1:t)} + z^{(1:t-1)}) + D_{\psi^*}(\theta^{(t+1)}||\theta^{(t)})$$

$$\psi^*(-z^{(1:T)}) = -\psi(w^{(1)}) - \sum_{t=1}^{T} < w^{(t)}, z^{(t)} > -D_{\psi^*}(-z^{(1:t)}|| - z^{(1:t-1)})$$

Substituting $\psi^*(-z^{(1:T)})$ back into the lower bound inequality and by rearranging terms, we are left with:

$$\psi(u) - u \cdot \theta^{(T+1)} \geq -\psi^*(\theta^{(T+1)})$$

$$< u, -z^{(1:T)} > -\psi(u) \leq \psi^*(-z^{(1:T)})$$

$$< u, -z^{(1:T)} > -\psi(u) \leq -\psi(w^{(1)}) - \sum_{t=1}^{T} < w^{(t)}, z^{(t)} > -D_{\psi^*}(-z^{(1:t)}|| - z^{(1:t-1)})$$

$$- < u, z^{(1:T)} > -\psi(u) \leq -\psi(w^{(1)}) - \sum_{t=1}^{T} < w^{(t)}, z^{(t)} > + \sum_{t=1}^{T} D_{\psi^*}(-z^{(1:t)}|| - z^{(1:t-1)})$$

$$\sum_{t=1}^{T} < w^{(t)}, z^{(t)} > - < u, z^{(1:T)} > \leq \psi(u) - \psi(w^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-z^{(1:t)}|| - z^{(1:t-1)})$$

Thus, the OMD regret bound is defined by the algorithm shown below. In the last line of the derviation on the right-hand-side, there are three terms. The first term is the regularization term, and the third term is the Bregman Divergence term under the convex conjugate of the regularization function.

$$\sum_{t=1}^{T} < w^{(t)}, z^{(t)} > - < u, z^{(1:T)} > \leq \psi(u) - \psi(w^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-z^{(1:t)}|| - z^{(1:t-1)})$$

$$R(u) \leq \underbrace{\psi(u)}_{\text{regularization}} - \psi(w^{(1)}) + \sum_{t=1}^{T} \underbrace{D_{\psi^*}(-z^{(1:t)}|| - z^{(1:t-1)})}_{\text{Bregman Divergence}}$$

References

[1] Wikipedia. Bregman divergence — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Bregman%20divergenceoldid=1070400527, 2022. [Online; accessed 17-February-2022].

[2] Wikipedia. Convex conjugate — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Convex%20conjugateoldid=1072291165, 2022. [Online; accessed 17-February-2022].