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

Online Mirror Descent

Lecturer: Kris Kitani *Scribes: Dakshit Agrawal, Lulu Ricketts*

1 Review

In the previous lectures, we studied general online optimization algorithms. Recall that online optimization consisted of the following steps:

1. Predict the weights (parameters).
2. Receive the loss function from nature.
3. Compute the loss using the predicted weights and given loss function.

In particular, we read about the following two algorithms:

1. Follow the Leader (FTL)
2. Follow the Regularized Leader (FTRL)

We also derived the regret bounds of these algorithms for linear and quadratic convex loss functions with weights in the convex space.

1.1 Follow the Leader (FTL)

The core principle of FTL is to choose those weights that have given the lowest loss till now (see Algo. 1).

Algorithm 1 Follow the Leader (FTL)

1: **for** $t = 1, 2, \ldots, T$ **do**
2: $\quad \mathbf{w}^{(t)} = \arg\min_{\mathbf{w} \in W} \sum_{i=1}^{t-1} f^{(i)}(\mathbf{w})$
3: $\quad$ RECEIVE $(f^{(t)} : W \to \mathbb{R})$
4: **end for**

We derived the regret bounds of FTL for the following convex functions:

1. Linear Loss ($f^{(t)}(w) = wz^{(t)}$):
   - Regret is $O(T)$
   - The algorithm can be unstable in certain cases leading to its failure.

2. Quadratic Loss ($f^{(t)}(w) = \frac{1}{2}\|w - z^{(t)}\|_2^2$):

   - Regret $\leq 4L^2\,(\log(T) + 1)$
   - No regret!
   - Stability in algorithm due to weight updates via averaging.

1.2 Follow the Regularized Leader (FTRL)

The core principle of FTRL is to ensure stability on weight prediction by adding a regularization term ψ to the weight prediction step (see Algo. 2). FTL is a specific case of FTRL in which $\psi = 0$.

Algorithm 2 Follow the Regularized Leader (FTRL)

1: **for** $t = 1, 2, \ldots, T$ **do**
2: $\quad \mathbf{w}^{(t)} = \arg\min_{\mathbf{w} \in W} \sum_{i=1}^{t-1} f^{(i)}(\mathbf{w}) + \psi(\mathbf{w})$
3: $\quad$ RECEIVE $(f^{(t)} : W \to \mathbb{R})$
4: **end for**

Taking a convex linear loss function like we did for FTL ($f^{(t)} = w \cdot z^{(t)}$), but now using a convex quadratic regularizer ($\psi(w) = \frac{1}{2\eta}\|w\|_2^2$), we get the regret bound of FTRL to be:

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

where

$$L = \max_{\mathbf{z}} \|\mathbf{z}\|_2$$
$$B = \max_{\mathbf{u} \in S} \|\mathbf{u}\|_2$$

As can be seen, the regret for a convex linear loss function is now **no regret** due to the quadratic regularizer.

Using online convex optimization, the stability of FTRL is generalized to:

- any sequence of Lipschitz loss functions (not just linear)
- other convex regularization functions (not just quadratic)

Thus, regularization can be used to ensure stability and no-regret properties.

2 Summary

We will first look at how FTRL can be interpreted as Online Mirror Descent (OMD) (Sec. 2.1. Afterwards, we will build our geometrical intuition and mathematical understanding of duality (Sec. 2.2) to finally calculate the regret bound for OMD (Sec. 2.3).

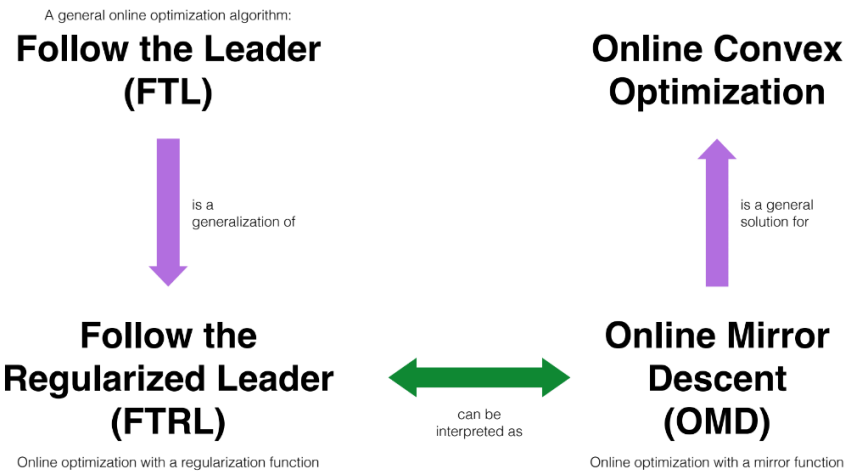

Figure 1: Overview of how topics taught till now connect to each other

2.1 Online Mirror Descent (OMD)

OMD provides another way to interpret FTRL when using a linear loss and convex regularizer. You might ask why we need this different perspective, since FTRL is stable and getting the job of online convex optimization done for us. The reason is two-fold:

- provides a unification of online learning algorithms
- more mathematical tools for regret analysis

In the subsequent subsections, we will generalize FTRL with linear loss ($f^{(i)}(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}^{(i)} \rangle$) to derive OMD.

2.1.1 Generalize FTRL linear loss parameter sum

The FTRL linear loss parameter sum is:

$$\sum_{i=1}^{t} f^{(i)}(\mathbf{w}) \tag{1}$$

The gradient of Eq. 1 with respect to the parameters $\mathbf{w}$ is:

$$\mathbf{z}^{(1:t)} = \sum_{i=1}^{t} \mathbf{z}^{(i)} \tag{2}$$

Let us substitute $\mathbf{z}^{(1:t)}$ in Eq. 2 with $\boldsymbol{\theta}^{(t+1)}$, where we call $\boldsymbol{\theta}$ a **parameter of the dual space**:

$$\boldsymbol{\theta}^{(t+1)} \triangleq -\mathbf{z}^{(1:t)} \tag{3}$$

Using Eq. 2 and Eq. 3, we get:

$$\boldsymbol{\theta}^{(t+1)} = -\sum_{i=1}^{t} \mathbf{z}^{(i)}$$
$$\boldsymbol{\theta}^{(t+1)} = -\sum_{i=1}^{t-1} \mathbf{z}^{(i)} - \mathbf{z}^{(t)}$$
$$\boldsymbol{\theta}^{(t+1)} = -\mathbf{z}^{(1:t-1)} - \mathbf{z}^{(t)}$$

Using Eq. 3:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mathbf{z}^{(t)} \tag{4}$$

Thus, we can represent Eq. 2 as an incremental sum of $\mathbf{z}$ as shown in Eq. 4.

2.1.2 Generalize FTRL prediction step

The FTRL prediction step is given by:

$$\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}} \sum_{i=1}^{t} f^{(i)}(\mathbf{w}) + \psi(\mathbf{w}) \tag{5}$$

Using a linear loss:

$$\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}} \sum_{i=1}^{t} \langle \mathbf{w}, \mathbf{z}^{(i)} \rangle + \psi(\mathbf{w})$$
$$\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}} \langle \mathbf{w}, \sum_{i=1}^{t} \mathbf{z}^{(i)} \rangle + \psi(\mathbf{w}) \quad \text{(sum of dot product is equal to dot product of sum)}$$
$$\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}} \langle \mathbf{w}, \mathbf{z}^{(1:t)} \rangle + \psi(\mathbf{w}) \quad \text{(Using Eq. 2)}$$
$$\mathbf{w}^{(t+1)} = \arg\max_{\mathbf{w}} -(\langle \mathbf{w}, \mathbf{z}^{(1:t)} \rangle + \psi(\mathbf{w})) \quad \text{(Converting min problem to max problem)}$$
$$\mathbf{w}^{(t+1)} = \arg\max_{\mathbf{w}} \langle \mathbf{w}, -\mathbf{z}^{(1:t)} \rangle - \psi(\mathbf{w})$$

Using Eq. 3 and defining $\mathbf{w}$ as a **parameter of the primal space**:

$$\mathbf{w}^{(t+1)} = \arg\max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta}^{(t+1)} \rangle - \psi(\mathbf{w}) \tag{6}$$

We can shorten Eq. 6 by defining a **mirror/linking function** $g : \boldsymbol{\theta} \to \mathbf{w}$ that maps from the dual space $\boldsymbol{\theta}$ to the primal space $\mathbf{w}$ such that:

$$\mathbf{w}^{(t+1)} \triangleq g(\boldsymbol{\theta}^{(t+1)}) \triangleq \arg\max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta}^{(t+1)} \rangle - \psi(\mathbf{w}) \tag{7}$$

2.1.3 Derive OMD

When we use linear loss and convex regularization, FTRL (Algo. 2) becomes OMD (see Algo. 3).

Algorithm 3 Online Mirror Descent (OMD)

1: **for** $t = 1, 2, \ldots, T$ **do**
2: RECEIVE $(f^{(t)} : W \to \mathbb{R})$
3: $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \mathbf{z}^{(t)}, \quad \mathbf{z} \in \partial f^{(t)}(\mathbf{w}^{(t)})$ ▷ Dual parameter update (Eq. 4)
4: $\mathbf{w}^{(t+1)} = g(\boldsymbol{\theta}^{(t+1)})$ ▷ Mirror projection (Eq. 7)
5: **end for**

OMD is therefore a generic algorithm for solving online convex optimization (linear loss and convex regularizer). Contrary to online gradient descent that optimizes directly in the primal space, OMD optimizes in the dual space and then mirrors in the primal space (see Fig. 2), thus getting the name.

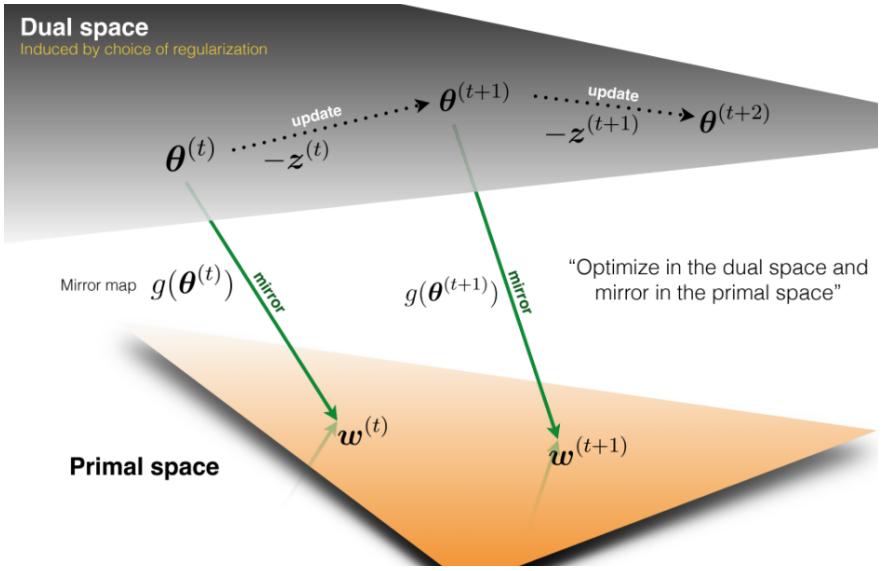

Figure 2: Online Mirror Descent visualization

Lastly, we can see that different choices of regularization leads to different mirror functions (Eq. 7), which allows us to take better advantage of the geometry of the solution space, consequently leading to different algorithms!

2.2 Duality

2.2.1 Convex Conjugate

To fully understand OMD, we must first understand convex conjugates, which are commonly known as a conjugate function, Fenchel dual/conjugate/transform, or Legrande transform. Here we assume

only smooth convex functions, however convex conjugates can take on other non smooth or convex forms. The equation for a convex conjugate is as follows:

$$\psi^*(\theta) = \max_{\mathbf{w}}(\langle \theta, w \rangle - \psi(w)) \tag{8}$$

To better understand the primal and dual spaces of OMD, it is important to know a function can be parameterized in two ways:

1. **Primal:** function / value parameterization: $\{\psi(w), w\}$
2. **Dual:** intercept / slope parameterization: $\{b(\theta), \theta\}$

As mentioned in a previous section, we compute a mirror function $g(\theta)$ for each time step, which we use to transform from the primal to dual space, and back again. The primal space is the one we are interested in and want to compute the parameters of. However, it may be easier to optimize in the dual space. OMD allows us the functionality to do so. See the appendix for more insight into conjugate functions.

The convex conjugate exhibits some interesting properties:

1. The gradient of the conjugate function is the optimal point on the function

$$\nabla_\theta \psi^*(\theta) = \frac{\partial \psi^*(\theta)}{\partial \theta} = w^* \tag{9}$$

2. Maximizer function

$$w^* = \arg\max \mathbf{w}(\langle \theta, w \rangle - \psi(w)) \tag{10}$$

3. Slope function

$$\nabla_w \psi(w) = \frac{\partial \psi^*(w)}{\partial w}|_{w=w^*} = \theta \tag{11}$$

4. Fenchel-Young Inequality (lower bound on conjugate function)

$$\psi^*(\theta) \geq (\langle w, \theta \rangle - \psi(w)) \tag{12}$$

2.2.2 Bregman Divergence

In the general case, the Bregman Divergence is an approximation error between two points, according to some proximity function ψ. What exactly this divergence is is dependent on how you define ψ. In the specific case of OMD, this proximity function is the regularizer. The definition of Bregman Divergence is as follows:

$$D_\psi(w||u) = \psi(w) - \psi(u) - \nabla\psi(u)^T(w-u) \tag{13}$$

2.3 Generic Regret Bound of OMD

Using concepts from Sec. 2.1 and Sec. 2.2, we will now derive the regret bound for OMD. Before we do so, we need to know the concept of telescoping:

Definition 1. *A telescoping series is a series whose partial sums eventually only have a fixed number of terms after cancellation, i.e., its general term is usually of the form* $t_n = a_n - a_{n+1}$.

This concept can be used to make an expression a telescoping series by adding and subtracting the same term such that the expression can now be expressed as a partial sum.

We will now prove below the generic upper regret bound of OMD. Recall that regret is defined as:

$$R(\mathbf{u}) = \sum_{t=1}^{T} \mathbf{w}^{(t)} \cdot \mathbf{z}^{(t)} - \mathbf{u} \cdot \mathbf{z}^{(t)} \tag{14}$$

Theorem 2. *(Regret bound of OMD)*

$$R(\boldsymbol{u}) \leq \psi(\boldsymbol{u}) - \psi(\boldsymbol{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\boldsymbol{z}^{(1:t)} \| - \boldsymbol{z}^{(1:t-1)}) \tag{15}$$

The regret of OMD is upper bounded by the total sum of the difference between regularization function and the sum of all of the Bregman Divergence under the convex conjugate of the regularization function.

Proof. The loss of an arbitrary vector $\mathbf{u}$ is given by:

$$\begin{aligned} &\psi(\mathbf{u}) + \sum_{t=1}^{T} \mathbf{u} \cdot \mathbf{z}^{(t)} \\ &= \psi(\mathbf{u}) - \mathbf{u} \cdot \boldsymbol{\theta}^{(T+1)} \quad \text{(Using Eq. 2 and Eq. 3)} \end{aligned}$$

Applying Fenchel-Young Inequality (Eq. 12):

$$\psi(\mathbf{u}) - \mathbf{u} \cdot \boldsymbol{\theta}^{(T+1)} \geq -\psi^*(\boldsymbol{\theta}^{(T+1)}) \tag{16}$$

Applying telescoping, we get:

$$\begin{aligned}-\psi^*(\boldsymbol{\theta}^{(T+1)}) &= -\psi^*(\boldsymbol{\theta}^{(T+1)}) - \psi^*(\boldsymbol{\theta}^{(T)}) + \psi^*(\boldsymbol{\theta}^{(T)}) - \cdots - \psi^*(\boldsymbol{\theta}^{(1)}) + \psi^*(\boldsymbol{\theta}^{(1)}) \\ &= -\psi^*(\boldsymbol{\theta}^{(1)}) - \sum_{t=1}^{T} \Big(\psi^*(\boldsymbol{\theta}^{(t+1)}) - \psi^*(\boldsymbol{\theta}^{(t)})\Big) \quad \text{(Sum over shifted time steps)}\end{aligned}$$

Using Eq. 13, we get:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = -\psi^*(\boldsymbol{\theta}^{(1)}) - \sum_{t=1}^{T} \Big(\nabla\psi^*(\boldsymbol{\theta}^{(t)}) \cdot (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) + D_{\psi^*}(\boldsymbol{\theta}^{(t+1)} || \boldsymbol{\theta}^{(t)})\Big) \tag{17}$$

The first term of Eq. 17 ($-\psi^*(\boldsymbol{\theta}^{(1)})$) can be written as $\psi(\mathbf{w}^{(1)})$ as shown below:

$$\begin{aligned}\psi^*(\boldsymbol{\theta}^{(1)}) &= \psi^*(\mathbf{z}^{(0)}) \quad \text{(Using Eq. 3)} \\ &= \psi^*(\mathbf{0}) \quad \text{(From initialization)} \\ &= \max_{\mathbf{w}} \{\langle \mathbf{w}, \mathbf{0} \rangle - \psi(\mathbf{w})\} \quad \text{(Using Eq. 8)} \\ &= \max_{\mathbf{w}} \{0 - \psi(\mathbf{w})\} \quad \text{(Compute dot product)} \\ &= -\min_{\mathbf{w}} \{\psi(\mathbf{w})\} \quad \text{(Convert to min problem)}\end{aligned}$$

At $t = 1$, the minimizer of cumulative loss is the primal iterate at step 1 ('the one-step look ahead cheater', discussed while deriving the intermediate regret bound of Algo. 2). Thus we get:

$$\psi^*(\boldsymbol{\theta}^{(1)}) = -\psi(\mathbf{w}^{(1)}) \tag{18}$$

Eq. 17 can be further reduced:

$$\begin{aligned}-\psi^*(\boldsymbol{\theta}^{(T+1)}) &= -\psi^*(\boldsymbol{\theta}^{(1)}) - \sum_{t=1}^{T} \Big(\nabla\psi^*(\boldsymbol{\theta}^{(t)}) \cdot (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) + D_{\psi^*}(\boldsymbol{\theta}^{(t+1)} || \boldsymbol{\theta}^{(t)})\Big) \quad \text{(From Eq. 17)} \\ &= \psi(\mathbf{w}^{(1)}) - \sum_{t=1}^{T} \Big(\nabla\psi^*(\boldsymbol{\theta}^{(t)}) \cdot (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) + D_{\psi^*}(\boldsymbol{\theta}^{(t+1)} || \boldsymbol{\theta}^{(t)})\Big) \quad \text{(Using Eq. 18)} \\ &= \psi(\mathbf{w}^{(1)}) - \sum_{t=1}^{T} \Big(\mathbf{w}^{(t)} \cdot (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) + D_{\psi^*}(\boldsymbol{\theta}^{(t+1)} || \boldsymbol{\theta}^{(t)})\Big) \quad \text{(Using Eq. 9)} \\ &= \psi(\mathbf{w}^{(1)}) - \sum_{t=1}^{T} \Big(\mathbf{w}^{(t)} \cdot (-\mathbf{z}^{(1:t)} + \mathbf{z}^{(1:t-1)}) + D_{\psi^*}(-\mathbf{z}^{(1:t)} || -\mathbf{z}^{(1:t-1)})\Big) \quad \text{(Using Eq. 3)} \\ &= \psi(\mathbf{w}^{(1)}) - \sum_{t=1}^{T} \Big(\langle \mathbf{w}^{(t)}, -\mathbf{z}^{(t)} \rangle + D_{\psi^*}(-\mathbf{z}^{(1:t)} || -\mathbf{z}^{(1:t-1)})\Big)\end{aligned}$$

Moving negative out of the sum:

$$-\psi^*(\boldsymbol{\theta}^{(T+1)}) = \psi(\mathbf{w}^{(1)}) + \sum_{t=1}^{T} \Big(\langle \mathbf{w}^{(t)}, \mathbf{z}^{(t)} \rangle - D_{\psi^*}(-\mathbf{z}^{(1:t)} || -\mathbf{z}^{(1:t-1)})\Big) \tag{19}$$

Combining Eq. 16 and Eq. 19, we get:

$$\psi(\mathbf{u}) - \mathbf{u}\cdot\boldsymbol{\theta}^{(T+1)} \geq \psi(\mathbf{w}^{(1)}) + \sum_{t=1}^{T}\Big(\langle \mathbf{w}^{(t)}, \mathbf{z}^{(t)}\rangle - D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)})\Big)$$

$$-\psi(\mathbf{u}) + \mathbf{u}\cdot\boldsymbol{\theta}^{(T+1)} \leq -\psi(\mathbf{w}^{(1)}) - \sum_{t=1}^{T}\Big(\langle \mathbf{w}^{(t)}, \mathbf{z}^{(t)}\rangle - D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)})\Big) \quad \text{(Multiplying by negative sign)}$$

$$-\psi(\mathbf{u}) + \langle \mathbf{u}, -\mathbf{z}^{(1:T)}\rangle \leq -\psi(\mathbf{w}^{(1)}) - \sum_{t=1}^{T}\Big(\langle \mathbf{w}^{(t)}, \mathbf{z}^{(t)}\rangle - D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)})\Big) \quad \text{(Using Eq. 3)}$$

$$-\psi(\mathbf{u}) + \langle \mathbf{u}, -\mathbf{z}^{(1:T)}\rangle \leq -\psi(\mathbf{w}^{(1)}) - \sum_{t=1}^{T}\langle \mathbf{w}^{(t)}, \mathbf{z}^{(t)}\rangle + \sum_{t=1}^{T} D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)}) \quad \text{(Multiply through sign)}$$

Rearranging terms:

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

$$\sum_{t=1}^{T}\langle \mathbf{w}^{(t)}, \mathbf{z}^{(t)}\rangle + \sum_{t=1}^{T}\langle \mathbf{u}, -\mathbf{z}^{(t)}\rangle \leq \psi(\mathbf{u}) - \psi(\mathbf{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)}) \quad \text{(Using Eq. 2)}$$

$$\sum_{t=1}^{T}\langle \mathbf{w}^{(t)}, \mathbf{z}^{(t)}\rangle - \langle \mathbf{u}, \mathbf{z}^{(t)}\rangle \leq \psi(\mathbf{u}) - \psi(\mathbf{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)}) \quad \text{(Combining sums)}$$

$$\sum_{t=1}^{T}\mathbf{w}^{(t)}\cdot\mathbf{z}^{(t)} - \mathbf{u}\cdot\mathbf{z}^{(t)} \leq \psi(\mathbf{u}) - \psi(\mathbf{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)}) \quad \text{(Changing notations)}$$

$$R(\mathbf{u}) \leq \psi(\mathbf{u}) - \psi(\mathbf{w}^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(-\mathbf{z}^{(1:t)}||-\mathbf{z}^{(1:t-1)}) \quad \text{(Using Eq. 14)}$$

We have thus proved the generic upper regret bound of OMD.

3 Appendix

3.1 Geometrical Interpretation of Conjugate Function

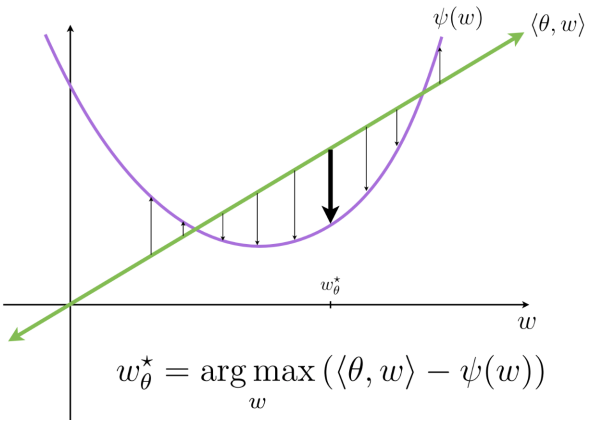

Figure 3: Conjugate Function Geometry

To get a better idea of conjugate functions, Figure (3) shows a geometrical interpretation of the conjugate function. In this, we have our primal function, a line $\langle \theta, w \rangle$ that passes through the origin with slope θ. The dual parameterization is $\psi(w)$ parameterized by w, in which we are trying to find the minimum w_θ^* of. Consider the difference between the primal and dual parameterizations. From the figure, we can see that the maximum difference actually occurs at the location where the tangent line meets $\psi(w)$, and also has a slope of θ. In other words, it is parallel to $\langle \theta, w \rangle$. The conjugate function at $\psi(\theta)$ is the intercept of the tangent line, which is therefore equivalent to the maximum difference of $\langle \theta, w \rangle - \psi(w)$.

3.2 Geometrical Interpretation of Bregman Divergence

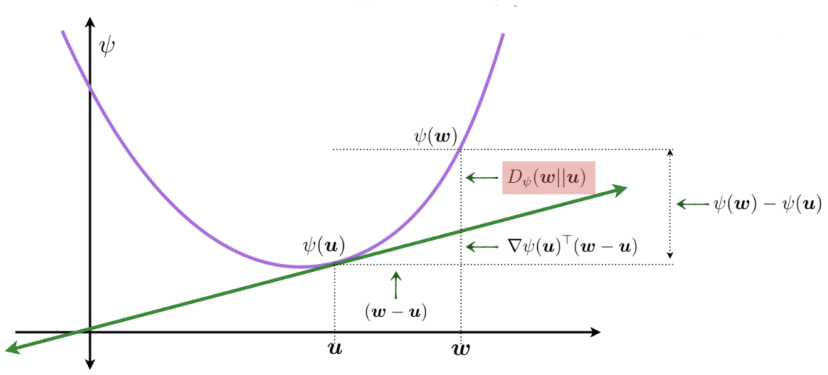

Figure 4: Bregman Divergence Geometry

Figure (4) shows a geometric interpretation of the Bregman Divergence that may be more intuitive to understand. Given two points u and v, we can calculate their difference as well as their difference

in values on the proximity function ψ, as shown on the bottom and right sides of the figure, respectively. As we can observe, the Bregman Divergence estimates the approximate error between the points by finding the difference between (1) the difference of their function values $\psi(w)$ - $\psi(u)$ and (2) the first order approximation of $\psi(w)$, $\nabla\psi(u)^T(w-u)$. This will come in handy as we form the regret bound for OMD.

References

[1] Wikipedia contributors. Bregman divergence – Wikipedia, the free encyclopedia, 2022. [Online; accessed 16-Feb-2022].

[2] Wikipedia contributors. Convex conjugate – Wikipedia, the free encyclopedia, 2022. [Online; accessed 16-Feb-2022].