Statistical Techniques in Robotics (16-831, S22) Lecture #09 (Wednesday, February 16)

Online Gradient Descent

Lecturer: Kris Kitani *Scribes: Seth Karten, Siva Kailas (Group 1)*

1 Review

In the last lectures, on the topic of Online Mirror Descent, we generalized Follow-the-Regularized-Leader (FTRL) with linear loss (FTRL-LinLoss) to derive the Online Mirror Descent (OMD). For a quick high level takeaway, we note that OMD does optimization in the dual space and "mirrors" the optimization step in the primal space. This is done through a mirror function $g(\theta)$. In this lecture, we will cover different perspectives on gradient descent. Then show that online gradient descent with a batch size of 1 is simply online mirror descent. Finally, we will provide a regret analysis of online gradient descent.

1.1 Online Mirror Descent

Recall that we defined $\theta^{t+1} \triangleq -z^{1:t}$ as the parameter of the dual space and denoted $z^{1:t} = \sum_{i=1}^{t} z^t$ as the sum of gradients. Thus, we can represent θ^{t+1} as shown below.

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

Furthermore, we defined w^{t+1} as the parameter of the primal space. This allowed to represent w with the equality $w^{t+1} = g(\theta^{t+1})$, where $g : \theta \to w$ is the mirror function or linking function. This allows us to derive the OMD algorithm as shown below.

Algorithm 1 Online Mirror Descent (Convex Set S, $g : \mathbf{R}^D \to S$)

1: **for** $t = 1, ..., T$ **do**
2: `RECEIVE`$(f^t : S \to \mathbf{R})$
3: $\theta^{t+1} = \theta^t - \eta z^t$, $z \in \partial f^t(w^t)$
4: $w^{t+1} \leftarrow g(\theta^{t+1})$
5: **end for**

1.2 Duality

1.2.1 Conjugate Function

Recall that the convex conjugate function $\psi^*(\theta)$ was defined as shown below.

$$\psi^*(\theta) = \max_w(\langle \theta, w \rangle - \psi(w))$$

We can define the derivative of the convex conjugate function $\psi^*(\theta)$ as shown below.

$$\nabla_\theta \psi^*(\theta) = \frac{\partial \psi^*(\theta)}{\partial \theta} = w^* = \arg\max_w(\langle \theta, w \rangle - \psi(w))$$

We can also define the Fenchel-Young inequality as shown below.

$$\psi^*(\theta) \geq (\langle w, \theta \rangle - \psi(w))$$

Finally, we can define the Bregman Divergence as shown below.

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

The Bregman Divergence characterizes the 'distance' between two points w and u according to the proximity function ψ. Note that in OMD, ψ is the regularization function.

1.3 OMD Analysis

We now summarize the regret analysis of the OMD algorithm.

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

2 Online Gradient Descent

2.1 Geometric Perspective

The idea of online gradient descent from the geometric perspective is to move in the direction opposite of the gradient.

Definition 1. *The gradient of a differentiable function, $f : \mathbb{R}^N \to \mathbb{R}$, is denoted by $\nabla f(w)$, which is defined at some vector w. We say our hypothesis is a vector w in a convex hypothesis class. $\nabla f(w)$ is a vector of partial derivatives $\nabla f(w) = \{\frac{\partial f(w)}{\partial w_1}, ..., \frac{\partial f(w)}{\partial w_N}\}$.*

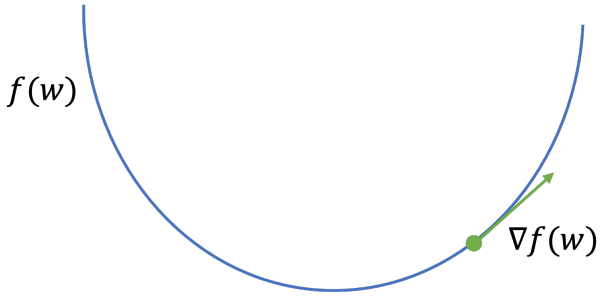

Figure 1: Geometric gradient intuition.

From the graph of our convex function (Fig. 1), we can intuitively see that if one wants to find the minima of f, one must move in the opposite direction of the gradient. Thus, we can define the algorithm for gradient descent as follows: In the above algorithm, we used our intuition about the geometry to define the update step in line 4. We move each weight in the direction of the negative gradient. The magnitude that we move is scaled by our learning rate ∇. Later, we will see how careful tuning of ∇ will affect the regret.

Algorithm 2 Gradient Descent(f)

1: $w^0 \leftarrow 0$
2: **for** $t = 1, ..., T$ **do**
3: `COMPUTE`$(\nabla f(w^{t-1}))$
4: $w^t \leftarrow w^{t-1} - \eta \nabla f(w^{t-1})$
5: **end for**

2.2 Linear Approximation with Regularization Perspective

In order to add some rigor to our intuition about the gradient update, let us explore gradient descent from the perspective of linear approximation with regularization. By taking a first-order

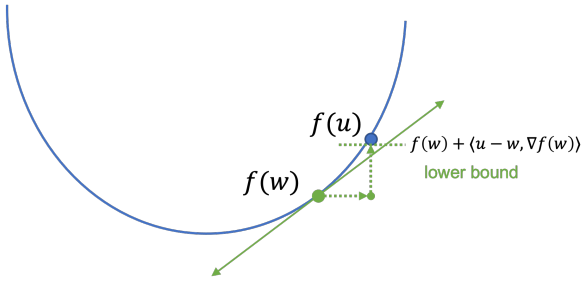

Figure 2: Showing the lower bound of the first-order Taylor series approximation.

Taylor series expansion of f around w, we can approximate the value at some point u, $f(u) \approx f(w) + \langle u - w, \nabla f(w) \rangle$.

Lemma 2. *When f is convex, the first-order Taylor series approximation is a lower bound.*

Given the preceding lemma, we know, $f(u) \geq f(w) + \langle u - w, \nabla f(w) \rangle$. See Fig. 2.

One may think that we can minimize this approximation, $\min_u \{f(w) + (u - w)^\intercal \nabla f(w)\}$, directly, but this will result in a solution at negative infinity.

Recall that linear approximations are only accurate for values 'close to' w (ϵ-neighborhood). We can introduce a regularization term to the objective function to constrain the expression and express closeness. One such constraint is the squared L2 norm, $\min_w ||w - w^t||_2^2$. Thus, our final objective function becomes a linear loss function with a quadratic regularization term:

$$w^{t+1} = \arg\min_w \frac{1}{2}||w - w^t||^2 + \eta\big(f(w^t) + \langle w - w^t, \nabla f(w^t) \rangle\big)$$

The first term in the objective function's purpose is regularization to keep the w values close to w^t. At the same time, the second term in the objective function looks to find the minimum w with respect to minimizing the approximation of f. The η, which we traditionally think of as our learning rate, is set up as the Lagrange multiplier in this dual objective optimization problem.

By taking the derivative of the objective function and setting it equal to 0, we can solve for w^{t+1}, deriving the gradient update step, $w^{t+1} = w^t - \eta \nabla f(w^t)$.

2.3 Isometric Quadratic Approximation Perspective

We know that online mirror descent can be used to solve functions with linear loss and quadratic regularization, so we need to analyze another perspective.

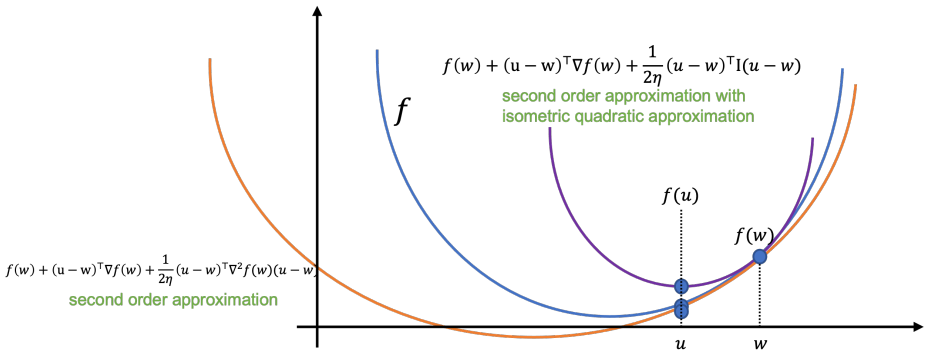

Figure 3: The isometric approximation is shown above, compared to the original function and the second-order approximation.

A second-order Taylor expansion yields,

$$f(u) \approx f(w) + (u-w)^\intercal \nabla f(w) + \frac{1}{2}(u-w)^\intercal \nabla^2 f(w)(u-w).$$

We can add a tunable scaling parameter to the second-order term and use an isometric quadratic approximation, i.e., unit scaling in all dimensions/directions, to get,

$$f(u) \approx f(w) + (u-w)^\intercal \nabla f(w) + \frac{1}{2\eta}(u-w)^\intercal \mathbf{I}(u-w).$$

By multiplying the RHS by η and rearranging the terms, we have derived the same objective function as before,

$$w^{t+1} = \arg\min_w \frac{1}{2}||w-w^t||^2 + \eta\big(f(w^t) + \langle w-w^t, \nabla f(w^t)\rangle\big)$$

Similarly, we can take the derivative of the objective function and set it equal to 0.

$$\frac{\partial}{\partial w}\Big\{\frac{1}{2}||w-w^t||^2 + \eta\Big(f(w^t) + \langle w-w^t, \nabla f(w^t)\rangle\Big)\Big\} = 0$$

$$\frac{1}{2}\Big(2w + 0 - 2w^t\Big) + \eta\Big(0 + \nabla f(w^t) - 0\Big) = 0$$

$$w - w^t + \eta\nabla f(w^t) = 0$$

Then solve for w to derive the gradient update step,

$$w = w^t - \eta\nabla f(w^t),$$

which we can use for $w^{t+1} = w^t - \eta\nabla f(w^t)$.

2.4 Summary

We have shown that we can find the parameter update,

$$w^{t+1} = w^t - \eta \nabla f(w^t),$$

as a result of:

1. a quadratic approximation of the loss function f
2. a linear loss function f with quadratic regularization
3. a quadratic loss function with linear constraints

Item 3 will be seen in the next lecture.

3 OGD/SGD vs. OMD

In this section, we will show how online gradient descent (OGD) may be equivalent to stochastic gradient descent (SGD) and how they are a special case of online mirror descent (OMD) [1].

3.1 Stochastic Gradient Descent

We can speed up online gradient descent through stochastic gradient descent. Stochastic gradient descent is a convex optimizer which does not require the exact gradient. The direction of the descent will be a random vector. SGD only requires that the expected value at each iteration will equal the gradient direction.

Algorithm 3 Stochastic Gradient Descent(f)

1: $w^1 \leftarrow 0$
2: $\eta > 0$
3: **for** $t = 1, ..., T$ **do**
4: $\quad z \sim D$
5: $\quad v^t = \nabla f_z(w^{t-1})$
6: $\quad w^t = w^{t-1} - \eta v^t$
7: **end for**

We detail the stochastic gradient descent algorithm above. In line 4, the algorithm samples from the data distribution, which can be a single sample or a mini-batch. In line 5, we can quickly compute the gradient of a single sample or mini-batch. Since the expectation is the true gradient, the algorithm still holds. One may recall that this looks exactly like the online learning algorithm without x, and this looks exactly like online convex optimization.

3.2 Method Connections

We can connect Online gradient descent (OGD) / stochastic gradient descent to be a special case of online mirror descent. Define the regularization function as, $\phi(w) = \frac{1}{2\eta}||w||_2^2$, which is quadratic. Define the loss function as, $f(w) = \langle w, \theta \rangle$, which is linear. Then the prediction rule becomes,

$$w^t + 1 = \arg\min_w \langle w, -\theta^{t+1} \rangle + \phi(w).$$

Substituting in our regularization function, we have,

$$w^t + 1 = \arg\min_w \langle w, -\theta^{t+1} \rangle + \frac{1}{2\eta}||w||_2^2.$$

For this combination, we can derive the analytical solution.

$$w^t + 1 = \arg\min_w \langle w, -\theta^{t+1} \rangle + \frac{1}{2\eta}\sqrt{\sum_n w_n^2}^2$$

This is equivalent to,

$$w^t + 1 = \arg\min_w \langle w, -\theta^{t+1} \rangle + \frac{1}{2\eta}\sum_n w_n^2.$$

Define, $\mathbf{L} = \langle w, -\theta^{t+1} \rangle + \frac{1}{2\eta}\sum_n w_n^2$. We can take the derivative with respect to w_n to find the minima,

$$\frac{\partial \mathbf{L}}{\partial w_n} = -\theta_n + \frac{1}{2\eta} 2 w_n = 0.$$

Thus, we have found the optimal parameter, $w_n = \eta\theta_n$, which is also a mapping from the dual parameter to the primal parameter. That is, we have found the mirror function for OGD, $g(\theta) = \eta\theta$.

We know that there are two versions of the OGD mirror function. The first is,

$$g(\theta) = \Pi_{\theta \to S}(n\theta),$$

which is a projection to the convex set. See algorithm 5 line 2 for the dual parameter update and line 3 for the mirror projection with some explicit condition on the feasibility set via projection. The other is,

$$\Pi_{x \to S}(x) = \arg\min_y ||x - y||^2,$$

which, for example, is a euclidean projection to S. We found,

$$g(\theta) = \eta\theta,$$

which has no projection. This assumes a constrained range of sub-gradients, which will slightly change the regret analysis. See algorithm 4 line 2 for the dual parameter update and assumption on the magnitude of the sub-gradient and line 3 for the mirror projection.

Thus, we have shown that OGD is OMD with a linear loss and quadratic regularization (or quadratic loss and linear constraints).

Algorithm 4 Online Sub-Gradient Descent(η)

1: **for** $t = 1, ..., T$ **do**
2: $\theta^{t+1} = \theta^t - z^t$, $z^t \in \partial f^t(w^t)$
3: $w^{t+1} = \eta\theta^{t+1}$
4: **end for**

Algorithm 5 Online Proj Sub-Gradient Descent(η)

1: **for** $t = 1, ..., T$ **do**
2: $\theta^{t+1} = \theta^t - z^t$, $z^t \in \partial f^t(w^t)$
3: $w^{t+1} = \Pi_{\theta \to S}\eta\theta^{t+1}$
4: **end for**

3.2.1 Quick Aside to Other Algorithms

We can quickly see that this is similar to the perceptron algorithm with the following derivation. From

$$w^{t+1} = \eta\theta^{t+1},$$

we expand the mirror function, $w^{t+1} = -\eta \sum_{i=1}^{t} z^i$. Separating the sum, we have, $w^{t+1} = -\eta(z^t + \sum_{i=1}^{t-1} z^i)$. We can convert this to shorthand notation, $w^{t+1} = -\eta(z^t + \theta^t)$. By definition of the weight vector, $w^{t+1} = -\eta(z^t - \frac{1}{\eta}w^t)$. Then finally simplifying, $w^{t+1} = w^t - \eta z^t$.

4 Online Gradient Descent Analysis

To begin the analysis of OGD, we first start with the general regret bound derived for OMD as shown below.

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

The regularization function is defined as $\psi(w) = \frac{1}{2}||w||_2^2$. From this, we can derive that the convex conjugate can be defined as $\psi^*(\theta) = \frac{1}{2}||\theta||_2^2$. To do so, we first start with the convex conjugate definition and rearrange the terms as shown below.

$$\psi^*(\theta) = \max_w(\langle\theta, w\rangle - \psi(w)) = -\min_w(\psi(w) - \langle\theta, w\rangle)$$

$$\psi^*(\theta) = -\min_w(\frac{1}{2}||w||_2^2 - \langle\theta, w\rangle)$$

Next, we solve for the minima by finding for what w the first-order optimality conditions are satisfied, as shown below.

$$\frac{\partial}{\partial w}\{\frac{1}{2}||w||_2^2 - \langle\theta, w\rangle)\} = 0$$

$$w - \theta = 0$$

$$w = \theta$$

Now, we revisit the original inequality $\psi^*(\theta) = -\min_w(\frac{1}{2}||w||_2^2 - \langle\theta, w\rangle)$ and rearrange the terms to derive the desired result as shown below.

$$\psi^*(\theta) = -\min_w(\frac{1}{2}||w||_2^2 - \langle\theta, w\rangle)$$

$$-\psi^*(\theta) = (\frac{1}{2}||\theta||_2^2 - \langle\theta, \theta\rangle)$$

$$-\psi^*(\theta) = (\frac{1}{2}||\theta||_2^2 - ||\theta||_2^2)$$

$$-\psi^*(\theta) = -\frac{1}{2}||\theta||_2^2$$

$$\psi^*(\theta) = \frac{1}{2}||\theta||_2^2$$

Now, we can return to the general regret bound derived for OMD and replace the Bregman Divergence term with its corresponding definition, as shown below.

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

Next, we apply the definition of ψ and the derived ψ^* from above as shown below.

$$R(u) \leq \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^1||_2^2 + \sum_{t=1}^{T}\frac{1}{2\eta}||\theta^{t+1}||_2^2 - \frac{1}{2\eta}||\theta^t||_2^2 - \nabla\frac{1}{2\eta}||\theta^t||_2^2(\theta^{t+1} - \theta^t)$$

Next, we note that $\nabla\frac{1}{2\eta}||\theta^t||_2^2 = \frac{1}{\eta}\theta^t$, so we can replace the term $\nabla\frac{1}{2\eta}||\theta^t||_2^2$ as shown below

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

Next, we note that $\frac{1}{2\eta}||\theta^{t+1}||_2^2 - \frac{1}{2\eta}||\theta^t||_2^2 - \frac{1}{\eta}\theta^t(\theta^{t+1} - \theta^t) = \frac{1}{2\eta}||\theta^{t+1} - \theta^t||_2^2$ by completing the square. Thus, we can replace the terms within the summation with the completed square term, as shown below.

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

Next, using the incremental update definition $\theta^{t+1} = \theta^t - \eta z^t$, we can replace θ^{t+1} in the expression as shown below.

$$R(u) \leq \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^1||_2^2 + \sum_{t=1}^{T}\frac{1}{2\eta}||\theta^t - \eta z^t - \theta^t||_2^2 = \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^1||_2^2 + \sum_{t=1}^{T}\frac{1}{2\eta}||-\eta z^t||_2^2$$

Since $-\eta$ is a constant, it can be moved to the outside of the norm as $|\eta|^2$ as shown below.

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

Finally, since $\frac{1}{2\eta}||w^1||_2^2$ is always positive, we can remove it to derive a larger upper bound as shown below.

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

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

We define $D = \max ||u||_2 : u \in S$ as the maximum magnitude of the primal parameter. Furthermore, we define $G = \max ||z||_2 : z \in \partial f(w)$ as the maximum magnitude of the sub-gradient. We can use these to express the above bound as shown below.

$$R(u) \leq \frac{1}{2\eta}||u||_2^2 + \sum_{t=1}^{T}\frac{\eta}{2}||z^t||_2^2 \leq \frac{D^2}{2\eta} + \frac{\eta}{2}TG^2$$

This is a general regret bound, but we can determine the regret bound for the optimal η with respect to the above general regret bound. To do this, we compute the first order optimality conditions with respect to η as shown below.

$$\frac{\partial}{\partial\eta}(\frac{D^2}{2\eta} + \frac{\eta}{2}TG^2) = 0$$

$$-\frac{D^2}{2\eta^2} + \frac{TG^2}{2} = 0$$

$$\frac{D^2}{2\eta^2} = \frac{TG^2}{2}$$

$$D^2 = \eta^2 TG^2$$

$$\frac{D^2}{TG^2} = \eta^2$$

$$\eta = \sqrt{\frac{D^2}{TG^2}} = \frac{D}{G\sqrt{T}}$$

Note that since $\eta > 0$, we do not consider the negative root. Using this value of $\eta = \frac{D}{G\sqrt{T}}$, we can compute the regret bound as shown below.

$$R(u) \leq \frac{D^2}{2\eta} + \frac{\eta}{2}TG^2$$

$$R(u) \leq \frac{D^2}{2\frac{D}{G\sqrt{T}}} + \frac{\frac{D}{G\sqrt{T}}}{2}TG^2$$

$$R(u) \leq DG\sqrt{T}$$

Thus, we have derived the regret bound for OGD as $R_{OGD} \leq DG\sqrt{T}$ [2].

5 Online Normalized Exponentiated Gradient Descent

We note that the regularization function chosen above is not the only one we can choose. Suppose we choose to define the regularization function ψ as negative entropy with respect to w. That is, we can define $\psi(w)$ as shown below.

$$\psi(w) = \sum_{k=1}^{K}(w_k \log w_k) : w \in \mathbf{S}^K$$

Here, $\mathbf{S}^k$ is a K-simplex constraint, and negative entropy is convex. We can define the loss function as $f(w) = \langle w, \theta \rangle$, which is linear. The prediction rule can be represented as shown below.

$$w^{t+1} = \arg\min_{w} \langle w, -\theta^{t+1} \rangle + \psi(w)$$

Replacing ψ as the negative entropy with respect to w and ensuring $w \in \mathbf{S}^K$ yields the following prediction rule.

$$w^{t+1} = \arg\min_{w \in \mathbf{S}^K} \langle w, -\theta^{t+1} \rangle + \sum_{k=1}^{K}(w_k \log w_k)$$

We can migrate the simplex constraint to the objective to formulate a constrained Lagrange optimization function as shown below.

$$w^{t+1} = \arg\min_{w} \langle w, -\theta^{t+1} \rangle + \sum_{k=1}^{K}(w_k \log w_k) + \lambda(1 - \sum_{k=1}^{K} w_k)$$

$$\mathcal{L} = \langle w, -\theta^{t+1} \rangle + \frac{1}{\eta}\sum_{k=1}^{K}(w_k \log w_k) + \lambda(1 - \sum_{k=1}^{K} w_k)$$

We can solve the above constrained Lagrangian optimization problem by computing the first order necessary conditions for optimality, as shown below.

$$\frac{\partial \mathcal{L}}{\partial w_n} = -\theta_n + \frac{1}{\eta}(1 + \log w_n) - \lambda = 0$$

$$\frac{1}{\eta} \log w_n = \theta_n - \frac{1}{\eta} + \lambda$$

$$w_n = e^{\eta\theta_n - (1 - \eta\lambda)} = \frac{e^{\eta\theta_n}}{e^{(1-\eta\lambda)}}$$

We can use the following mirror function,

$$g(\theta) = \frac{\exp(\eta\theta)}{\sum_{n'} \exp(\eta\theta_{n'})}$$

Online Norm-Exp Gradient Descent (ONEGD) can be shown to derive the same update as the weighted majority algorithm. From

$$w_n^{t+1} = \frac{\exp(\eta\theta_n^{t+1})}{\sum_n' \exp(\theta_{n'})}$$

Algorithm 6 Online Norm-Exp_GD(η)

for $t = 1, ..., T$ **do**
  $\theta^{t+1} = \theta^t = \eta z^t$, $z^t \in \partial f^t(w^t)$
  $w^{t+1} \propto \exp(\eta\theta^{t+1})$
end for

with some algebraic tricks, it is equivalent to,

$$w_n^{t+1} = \frac{w_n^t \exp(-\eta z_n^t)}{\sum_k w_k^t \exp(-\eta z_k^t)}$$

Thus, we have,

$$w_n^{t+1} \propto w_n^t \exp(-\eta z_n^t)$$

The Hedge algorithm is an unnormalized exponentiated gradient descent algorithm.

Algorithm 7 Hedge Algorithm(β)

$w^1 \leftarrow \{w_n^1 = 1\}_{n=1}^N$
for $t = 1, ..., T$ **do**
  RECEIVE($x^t \in \{-1, 1\}$)
  $i \sim$ MULTINOMIAL(w^t/Φ^t)
  $\hat{y}^t = h_i(x^t)$
  RECEIVE($y^t \in \{-1, 1\}$)
  $w_n^{t+1} = w_n^t e^{\beta(1[y^t \neq h_n(x^t)])}$ $\forall n$
end for

References

[1] B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 525–533. JMLR Workshop and Conference Proceedings, 2011.

[2] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.