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

Online Gradient Descent

Lecturer: Kris Kitani *Scribes: Young Woo Kim, Daphne Chen*

1 Review

In the previous lecture we covered Online Mirror Descent (OMD). This is an extension to the content about convex optimization, including the Follow the Regularized Leader (FTRL) algorithm. FTRL can be interpreted as a form of OMD, thus in this lecture we will cover how OMD and OGD are connected to various online learning and optimization problems.

1.1 Follow the Regularized Leader (FTRL)

Follow the Regularized Leader [1], similarly to OMD, is a class of online optimization algorithms; however, it uses a regularization function instead of a mirror function in order to stabilize the loss function. In general, the algorithm uses the following steps:

Algorithm 1 Follow the Regularized Leader (FTRL)

1: **function** FTRL(Convex set S)
2: **for** $t = 1, \cdots, T$ **do**
3: $\mathbf{w}^{(t)} = argmin_{\mathbf{w}} \Sigma_{i=1}^{t-1} f^{(i)}(\mathbf{w}) + \psi(\mathbf{w})$ ▷ prediction rule with regularization
4: RECEIVE $(f^{(t)} : S \to \mathbb{R})$ ▷ one step loss
5: **end for**

Where the loss function is linear, and uses quadratic regularization, as follows:

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

$$f^{(t)} = w \cdot z^{(t)} \tag{2}$$

Where Equation 1 is the quadratic regularization and 2 is the linear loss.

The regret bound of the FTRL algorithm is given by the equation below:

$$R^{(T)}(\mathbf{u}) \leq BL\sqrt{2T} \tag{3}$$

In summary, the FTRL algorithm displays the following properties:

1. ***Varying regret:*** regret varies depending on the choice of loss function
2. ***Regularized:*** FTRL uses regularization to ensure stability and no-regret property

1.2 Online Mirror Descent (OMD)

We previously covered two forms of online optimization – FTRL and OMD. Online Mirror Descent uses a mirror function. FTRL and OMD can be interpreted as each other, but provide a different set of tools for online learning and regret analysis. OMD acts as a mathematical framework for connecting concepts between online learning and convex optimization.

The basis for the OMD algorithm are the steps shown below:

Algorithm 2 Online Mirror Descent (OMD)

function OMD(Convex set S, $g : \mathbb{R}^D \to S$)
for $t = 1, \cdots, T$ **do**
  Receive ($f^{(t)} : S \to \mathbb{R}$) ▷ mirror function
  $\theta^{(t+1)} = \theta^{(t)} - \eta z^{(t)}, z^{(t)} \in \partial f^{(t)}(w^{(t)})$
  $\mathbf{w}^{(t+1)} = g(\theta^{(t+1)})$ ▷ prediction rule
end for

The general idea for OMD is that it can adapt to the size of the gradient. The tradeoff, however, is that the algorithm may perform poorly when it diverges from the optimal parameters, as may happen when there is a significant change in the data during the online learning process, a problem which FTRL mitigates.

2 Summary

2.1 Gradient Descent

Gradient descent is the standard method we use to minimize differentiable convex functions. It can be used to minimize nonconvex functions as well, but it may converge to a local minimum instead of a guaranteed global minimum, as it would for convex functions. The lecture discusses three perspectives to understanding gradient descent.

2.1.1 Geometric Perspective

The geometric perspective of gradient is the intuitive idea of taking the gradient of a function and stepping in the direction opposite of it. Over many steps, we can imagine sufficient steps would approach the minimum of the function. This intuition is described in Algorithm 1

Algorithm 3 Gradient Descent (GD)

1: $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$ ▷ Weight initialization
2: **for** $t = 1, \cdots, T$ **do**
3: COMPUTE $(\nabla f(\mathbf{w}^{(t-1)}))$ ▷ Compute Gradient
4: $\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ ▷ Step opposite gradient direction
5: **end for**

Algorithm 1 provides the skeleton of gradient descent, but it is more intuitive than rigorous. For a function $f : \mathbb{R}^N \to \mathbb{R}$, the gradient is the vector $\nabla f(\mathbf{w}) = \{\frac{\partial f(\mathbf{w})}{w_1}, \cdots, \frac{\partial f(\mathbf{w})}{w_N}\}$.

2.1.2 Linear Approximation with Regularization

The second perspective on gradient descent recognizes that we can take a convex function we want to minimize, compute its first order Taylor series approximation at $\mathbf{w}$ and use it to lower bound the function at another point $\mathbf{u}$.

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle$$

Suppose we try to minimize this linear function directly; the solution would be at negative infinity, though this is obviously untrue of the original convex function. We must account for the fact that linear approximations are only accurate near the point $\mathbf{w}$ and must therefore be constrained by a regularizer that encourages closeness to $\mathbf{w}$. For an L2 regularizing term, the final objective function is as follows:

$$\mathbf{w}^{(t+1)} = \text{argmin}_w \frac{1}{2}||\mathbf{w} - \mathbf{w}^{(t)}||^2 + \eta\Big(f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle\Big)$$

2.1.3 Isometric Quadratic Approximation

The third perspective uses the second order approximation of the convex function f:

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

The second order gradient, however, can be expensive to compute, so we replace it with an identity matrix and introduce a scaling parameter to further approximate the convex function as such:

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

We can now see the second order term simplifies into the L2 regularization term from the linear approximation perspective.

$$\mathbf{w}^{(t+1)} = \text{argmin}_w \frac{1}{2}||\mathbf{w} - \mathbf{w}^{(t)}||^2 + \eta\Big(f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle\Big)$$

To solve this equation, we take the first order derivative of the expression we want the argmin of set it to zero. Doing so yields the familiar update equation for w, which we have seen in online perceptron, weighted majority, and follow the regularized leader.

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

2.2 Stochastic Gradient Descent

Generally, it is expensive to compute the gradient of the loss function f in the update equation because f is usually a sum over the losses of many data points. We can speed it up by performing the updates in batches, so the entire data set does not need to be evaluated for each update. To achieve maximal efficiency, we can set the batch size to 1 and update after each training example. This example is chosen at random, or stochastically. The new estimated gradient at each iteration will no longer equal the original gradient direction, but the expected gradient will. This algorithm is known as stochastic gradient descent and similar convergence bounds as gradient descent.

Algorithm 4 Stochastic Gradient Descent (SGD)

1: $\mathbf{w}^{(1)} \leftarrow \mathbf{0}$ ▷ Weight initialization
2: $\eta > 0$ ▷ Set learning rate
3: **for** $t = 1, \cdots, T$ **do**
4: $z \sim \mathcal{D}$ ▷ Mini-batch or Single Sample
5: $\mathbf{v}^{(t)} = \nabla f_z(\mathbf{w}^{(t-1)})$ ▷ Compute Faster Gradient
6: $\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta \nabla \mathbf{v}^{(t)}$ ▷ Step opposite gradient direction
7: **end for**

2.3 OGD/SGD as a Special Case of Mirror Descent (OMD)

Recall the Online Mirror Descent algorithm from Section 1.2. We will now observe how online gradient descent can be represented in the mirror descent framework. OMD updates parameters in the primal space by updating parameters in the dual space and mapping it to the primal parameters using a mirror function. OMD with the quadratic regularization function $\psi(w) = \frac{1}{2\eta}||w||_2^2$ and linear loss function $f(w) = \langle w, \theta \rangle$ results in the following prediction rule:

$$\begin{aligned} w^{(1+t)} &= \arg\min_w \langle w, -\theta^{(t+1)} \rangle + \frac{1}{2\eta}||w||_2^2 \\ &= \arg\min_w \langle w, -\theta^{(t+1)} \rangle + \frac{1}{2\eta}\sum_n w_n^2 \end{aligned}$$

We can now describe the loss function $\mathbb{L}$ as follows:

$$\mathbb{L} = \langle w, -\theta \rangle + \frac{1}{2\eta}\sum_n w_n^2$$

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

$$w_n = \eta \theta_n$$

This shows that the optimal parameter for w_n is $\eta\theta_n$, and that our mirror function for OGD is the trivial projection $g(\theta) = \eta\theta$. Note that this results in an update rule similar to the one in the perceptron algorithm. To summarize, online gradient descent is online mirror descent with a linear loss and quadratic regularization.

2.4 OGD Regret Analysis

In order to derive the regret bound for OGD, we will define the following equation:

$$R_{OGD} \leq DG\sqrt{T} \tag{4}$$

Where $D = max||\mathbf{u}||_2, u \in S$ (assumption on the magnitude of the primal parameter) and $G = max||\mathbf{z}||_2, z \in \partial f(\mathbf{w})$ (assumption on the magnitidue of the sub-gradient, related to dual parameter). We also need to recall the regularization function and the convex conjugate for OGD.

Let us assume that we use the L2 norm for the regularizer. Then the convex conjugate is also L2 norm, or:

$$\psi(w) = \frac{1}{2}||w||_2^2 \tag{5}$$

$$\psi^*(\theta) = \frac{1}{2}||\theta||_2^2 \tag{6}$$

From here, we can start from the general regret bound analysis for OMD to derive the regret bound for OGD.

$$R(u) \leq \psi(u) - \psi(w^{(1)}) + \sum_{t=1}^{T} D_{\psi^*}(\theta^{(t+1)}||\theta^{(t)}) \tag{7}$$

$$= \psi(u) - \psi(w^{(1)}) + \sum_{t=1}^{T} \psi^*(\theta^{(t+1)}) - \psi^*(\theta^{(t)}) - \nabla\psi^*(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) \tag{8}$$

Using the Bregman divergence under the L2 norm in Equations 7 and 8.

$$= \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^{(1)}||_2^2 + \sum_{t=1}^{T} \frac{1}{2\eta}||\theta^{(t+1)}||_2^2 - \frac{1}{2\eta}||\theta^{(t)}||_2^2 - \nabla\frac{1}{2\eta}||\theta^{(t)}||_2^2(\theta^{(t+1)} - \theta^{(t)}) \tag{9}$$

$$= \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^{(1)}||_2^2 + \sum_{t=1}^{T} \frac{1}{2\eta}||\theta^{(t+1)}||_2^2 - \frac{1}{2\eta}||\theta^{(t)}||_2^2 - \frac{1}{\eta}\theta^{(t)}(\theta^{(t+1)} - \theta^{(t)}) \tag{10}$$

In Step 9 computing the gradient, and in Step 10 completing the square.

$$= \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^{(1)}||_2^2 + \sum_{t=1}^{T} \frac{1}{2\eta}||\theta^{(t+1)} - \theta^{(t)}||_2^2 \tag{11}$$

$$= \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^{(1)}||_2^2 + \sum_{t=1}^{T} \frac{1}{2\eta}||\theta^{(t)} - \eta z^{(t)} - \theta^{(t)}||_2^2 \tag{12}$$

$$= \frac{1}{2\eta}||u||_2^2 - \frac{1}{2\eta}||w^{(1)}||_2^2 + \sum_{t=1}^{T} \frac{1}{2\eta}|| - \eta z^{(t)}||_2^2 \tag{13}$$

Where we plug in the definition of dual parameter in Step 11 and then subtract terms, finally removing the always positive term in Step 13.

Lastly, this results in the following bound:

$$\leq \frac{1}{2\eta}||u||_2^2 + \sum_{t=1}^{T} \frac{\eta}{2}||z^{(t)}||_2^2 \tag{14}$$

Thus we have derived the regret bound for OGD.

In order to derive the optimal step size η, assuming that T is known, we can take the derivative:

$$\frac{d}{d\eta}\{\frac{1}{\eta}D^2 + \frac{\eta}{2}G^2T\} = 0 \rightarrow \eta = \frac{D}{G\sqrt{T}} \tag{15}$$

And subsequently plug back in to the equation for OGD regret bound, in order to get the result:

$$R_{OGD}(U) \leq \frac{D^2}{2\eta} + \frac{\eta}{2}T^2G \tag{16}$$

$$= DG\sqrt{T} \tag{17}$$

2.5 Online Normalized Exponentiated Gradient Descent (ONEGD)

We have noted that different choices of regularization can result in different mirror functions and different algorithms. Another such algorithm is normalized exponentiated gradient descent. Instead of the quadratic regularization term from OGD, we now use the negative entropy regularization and the linear loss function.

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

$$f(w) = \langle w, \theta \rangle$$

This yields the following prediction rule:

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

We then add the simplex constraint to this objective:

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

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

By solving for the minimum of this loss function, we get the following w and mirror function

$$w_n = \frac{\exp(\eta\theta_k}{\exp(1 - \eta\lambda)}$$

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

Algorithm 5 Online Normalized Exponentiated Gradient Descent (ONEGD)

1: **function**ONLINE NORM-EXP-GD(η)
2: **for** $t = 1, \cdots, T$ **do**
3: $\quad \theta^{(t+1)} = \theta^{(t)} - \eta z^{(t)}, z^{(t)} \in \partial f^{(t)}(w^{(t)})$ ▷ Dual parameter update
4: $\quad \mathbf{w}^{(t+1)} \propto -\exp\left(\eta\theta^{(t+1)}\right)$ ▷ Mirror projection
5: **end for**

3 Conclusion

In this lecture we covered the topic of Online Gradient Descent. We learned about how this relates to the prior lectures on Online Mirror Descent and other forms of online learning problems. Delving into the topic of Online Gradient Descent, we covered the necessary background on gradient descent from three different perspectives, including the geometric perspective, using linear approximation with regularization, and isometric quadratic approximation. Subsequently, we algorithmically studied stochastic gradient descent, noting the similar convergence bounds between GD and SGD. Relating this back to earlier lectures, we analyzed how OGD and SGD can be described as special cases of OMD. We concluded by performing a regret bound analysis for OGD, followed by one last method, online normalized exponentiated gradient descent.

References

[1] H. B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *JMLR: WCP 15*, volume 15, 2011.

4 Appendix

4.1 Duality

This section will cover mathematical concepts that enable us to analyze OMD, namely conjugate functions and the Bregman divergence.

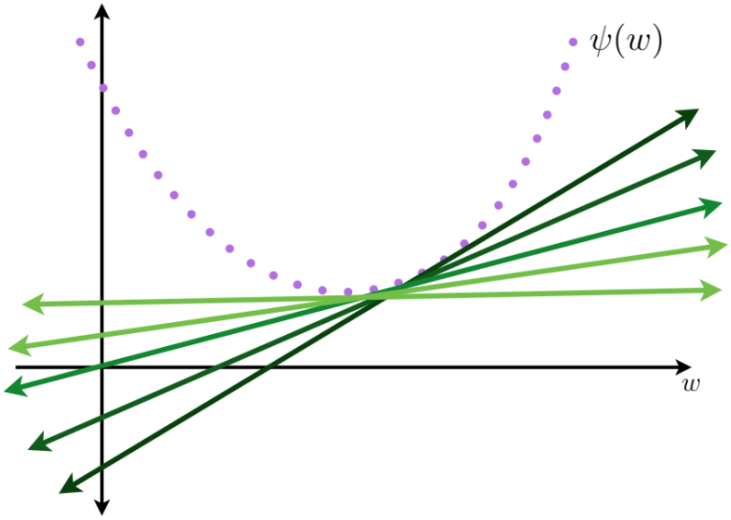
4.1.1 Conjugate Functions

A convex conjugate function, also known as a Fenchel conjugate, is a generalization of the Legendre transform. For the purpose of this course, we assume a smooth convex function, but the Legendre generalization can also be applied to non-smooth and non-convex functions.

The base formula for a convex conjugate function is as follows in Equation 18:

$$\psi^*(\theta) = \max_w(\langle \theta, w \rangle - \psi(w)) \tag{18}$$

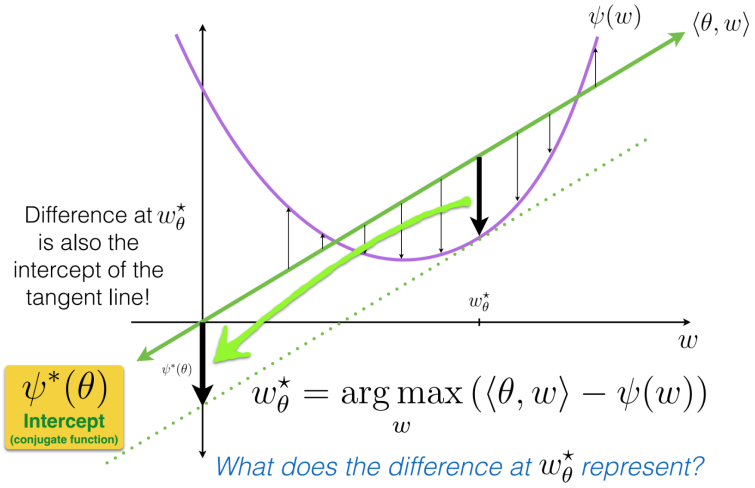
Next, we will analyze the geometry of the conjugate function. If we use intercept-slope parametrization (e.g. $\{\psi(w), w\}$, purple curve in below figure) with the function-value parametrization (e.g. $\{b(\theta), \theta\}$, green lines in below figure), as shown in the figure below (referenced from the class notes), then we can get the dual curve (or duality) that consists of a curve C and a set of lines tangent to C.



Given the function $\psi(w)$, the value θ is the derivative of ψ at $w = w^*$, thus the equation for the tangent line is $y = \langle \theta, w \rangle - b$. Using this formula, the "slope" or vertical distance between the y-intercept b and the tangent point is given as $\langle \theta, w \rangle$, or $\psi(w^*) + b$ using the curve in the figure. Thus the equation is updated to the following:

$$\langle \theta, w^* \rangle = \psi(w^*) + b(\theta) \tag{19}$$
$$-b(\theta) = -\langle \theta, w^*(\theta) \rangle + \psi(w^*(\theta)) \tag{20}$$

Using the geometry described above, we can now show how this holds for the conjugate function. The maximum distance between two functions, given in Equation 18, is equal to the intercept of the tangent line, shown visually in the below figure (referenced from the class notes).



The derivative of the convex conjugate function is given as:

$$\nabla_\theta \psi^*(\theta) = arg \max_w (\langle \theta, w \rangle - \psi(w)) \tag{21}$$

The Fenchel-Young inequality is:

$$\psi^*(\theta) \geq (\langle \theta, w \rangle - \psi(w)) \tag{22}$$

4.1.2 Bregman Divergence

The Bregman Divergence is represented in Equation 23:

$$D_\psi(w||u) = \psi(w) - \psi(u) - \nabla\psi(u)^T(w - u) \tag{23}$$

Where the distance D is defined between two points with a proximity function ψ. In the context of OMD, this means the regularization function.