CSC 588: Machine learning theory **Spring 2021**

Lecture 19: Analysis of online gradient descent; Online mirror descent: basic definitions

Lecturer: Chicheng Zhang *Scribe: Caleb Dahlke*

(Thursday, March 25.)

1 Online optimization

1.1 Online (sub)gradient descent algorithm

Initialize $w_1 \in \Omega$ and parameter η.
For $t = 1, 2, \ldots, T$:

- choose w_t
- Receive loss function f_t, suffer loss $f_t(w_t)$
- Set $g_t \in \partial f_t(w_t)$
- Update:
  $w'_{t+1} \leftarrow w_t - \eta g_t \qquad (\eta > 0)$
  $w_{t+1} \leftarrow \Pi_\Omega(w'_{t+1}) = \text{argmin}_{w\in\Omega} \|w - w'_{t+1}\|_2$

Remark 1.

$$w_{t+1} = \underset{w\in\Omega}{\text{argmin}}\, ||w - w_t + \eta g_t||_2^2 = \underset{w\in\Omega}{\text{argmin}} \langle w, \eta g_t\rangle + \frac{1}{2}||w - w_t||_2^2$$

Here $\langle w, \eta g_t\rangle$ can be seen as the correctiveness and $\frac{1}{2}||w - w_t||_2^2$ is seen as the conservativeness. See (Kivinen & Wamuth '97) for more information.

1.2 OGD Guarantees:

Theorem 2. *OGD w/ initializer w_1 and step size $\eta > 0$ grantees $\forall u \in \Omega$*

$$R_T(u) \leq \frac{||u - w_1||_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} ||g_t||_2^2$$

Moreover, if Ω has ℓ_2-diameter B ($\forall u, v \in \Omega$, $||u - v||_2 \leq B$) and $||g_t||_2 \leq \rho$ (which happens if all f_t's are ρ-Lipshitz) then

$$R_T(\Omega) \leq \frac{B^2}{2\eta} + \frac{\eta}{2}T\rho^2$$

Corollary 3. *Under the above setting, $\ell(w, z)$ is ρ-Lipshitz w.r.t. w, OGD with $f_t(w) = \ell(w, z_t)$ for i.i.d. $z_1, \ldots, z_T \sim D$ garuntees that $\overline{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$*

1. $\eta = \frac{B}{\rho}\sqrt{\frac{1}{T}} \Rightarrow \mathbb{E}[L_D(\overline{w}_T)] \leq \min_{w\in\Omega} L_D(w) + \frac{B\rho}{\sqrt{T}}$
2. $\eta = \frac{1}{\rho}\sqrt{\frac{1}{T}}, \Omega = \mathbb{R}^d, w_1 = 0 \Rightarrow \mathbb{E}[L_D(\overline{w}_T)] \leq L_D(w^*) + \frac{(||w^*||^2 + 1)\rho}{\sqrt{T}} \qquad \forall w^* \in \mathbb{R}^d$

Proof. *of Corollary*
Last time, we showed a high probability upper bound on

$$L_D(\overline{w}_T) \leq L_D(w^*) + \frac{R_t(\omega^*)}{T} + \text{Concentration}$$

From the proof, we have the online regret garuntee of the following form

$$\frac{1}{T}\sum_{t=1}^{T} \ell_t(w_t, z_t) - \frac{1}{T}\sum_{t=1}^{T} \ell_t(w^*, z_t) \leq R_T(w^*)$$

Take the expectation of the LHS

$$\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} \ell_t(w_t, z_t) - \frac{1}{T}\sum_{t=1}^{T} \ell_t(w^*, z_t)] = \mathbb{E}[\sum_{t=1}^{T} L_D(w_T)] - TL_D(w^*)$$

Then using the expectation upper bound and Jensen's inequality as before, as well a the given regret grantees, then one can prove the corollary. This is left out of lecture.

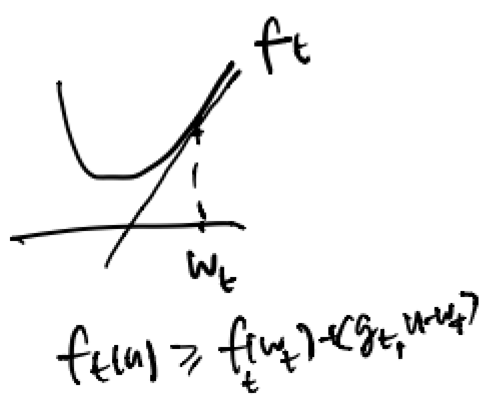
Chicheng notes: see my newly added Mar 23's scribe note, Remark 2, if the above is unclear. □

Now let us come back to the Online Gradient Decent Grantees and prove Theorem 2

Proof. **Step 1:** "linearization"
To start we, know

$$R_T(u) = \sum_{t=1}^{T} (f_t(w_t) - f_t(u))$$

we will bound $f_t(u)$ from below using the linearization shown in the following image



This means that we have the bound

$$R_T(u) = \sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq \sum_{t=1}^{T} \langle g_t, w_t - u \rangle$$

Step 2: "use optimality condition on w_{t+1}"
First order optimallity condition:
f is convex and differentiable in convex domain Ω. Call $x^* = \text{argmin}_{x\in\Omega} f(x)$. Then we have two cases

1. x^* is in the interior of Ω, then $\nabla f(x^*) = 0$ (if it weren't, we could walk in the direction of negative gradient to decrease the objective function, but this is assumed minimum)

2. x^* is in the boundary of Ω, we need $\forall y \in \Omega \ \langle \nabla f(x^*), y - x^* \rangle \geq 0$. Below is an illustration showing that with this condition, moving anywhere along the negative gradient would push us out of Ω

We can combine the two cases for the final result

$$x^* = \underset{x \in \Omega}{\operatorname{argmin}} f(x) \Leftrightarrow \forall y \in \Omega, \langle \nabla f(x^*), y - x^* \rangle \geq 0$$

The proof of this statement is omitted, but the outline is below

Idea of proof:

($\Rightarrow$) if $\exists y \ \langle \nabla f(x^*), y - x^* \rangle < 0$

$$f(x^* + \alpha(y - x^*)) = f(x^*) + \alpha \langle \nabla f(x^*), y - x^* \rangle + o(\alpha) < f(x^*) \text{ (for small } \alpha > 0)$$

($\Leftarrow$) $\forall y$

$$f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle \geq 0$$

The details to this need to be filled out.

Now lets apply this optimallity condition to the OGD, recall

$$w_{t+1} = \underset{w \in \Omega}{\operatorname{argmin}} \langle \eta g_t, w \rangle + \frac{1}{2} ||w - w_t||^2$$

First order optimallity

$$\langle \eta g_t + w_{t+1} - w_t, u - w_{t+1} \rangle \geq 0 \forall u \in \Omega$$

Now rewriting the above, we get

$$\langle g_t, w_{t+1} - u \rangle \leq \frac{1}{\eta} \langle w_{t+1} - w_t, u - w_{t+1} \rangle$$

We can now use the fact that $\langle a, b \rangle = \frac{1}{2} \left(||a + b||^2 - ||a|| - ||b|| \right)$

$$\langle g_t, w_{t+1} - u \rangle \leq \frac{1}{\eta} \langle w_{t+1} - w_t, u - w_{t+1} \rangle = \frac{1}{2\eta} \left(||u - w_t||^2 - ||u - w_{t+1}||^2 - ||w_{t+1} - w_t||^2 \right)$$

Step 3: "Bounding $\langle g_t, w_t - u \rangle$"

$$\langle g_t, w_t - u \rangle = \langle g_t, w_{t+1} - u \rangle + \langle g_t, w_t - w_{t+1} \rangle$$

We will now use Cauchy-Schwarz on the second term $\langle g_t, w_t - w_{t+1} \rangle \leq ||g_t|| ||w_t - w_{t+1}||$ and then we can use the geometric mean of these two numbers, that is $||g_t|| ||w_t - w_{t+1}|| = \eta 2 ||g_t||_2^2 + \frac{1}{2\eta} ||w_t - w_{t+1}||_2^2$.

$$\langle g_t, w_{t+1} - u \rangle + \langle g_t, w_t - w_{t+1} \rangle \leq \langle g_t, w_{t+1} - u \rangle + \eta 2 ||g_t||_2^2 + \frac{1}{2\eta} ||w_t - w_{t+1}||_2^2$$

Using the upper bound we developed in the previous step

$$\langle g_t, w_{t+1} - u\rangle + \eta_2 ||g_t||_2^2 + \frac{1}{2\eta}||w_t - w_{t+1}||_2^2 \leq \frac{\eta}{2}||g_t||_2^2 + \frac{1}{2\eta}\left(||u - w_t||^2 - ||u - w_{t+1}||^2\right)$$

Combining these parts together, we get

$$\langle g_t, w_t - u\rangle \leq \frac{\eta}{2}||g_t||_2^2 + \frac{1}{2\eta}\left(||u - w_t||^2 - ||u - w_{t+1}||^2\right)$$

This can be interpreted as if we have a large instantaneous regret then the iterate will be closer to the comparitor.

Step 4: "sum over t"

$$\sum_{t=1}^{T}\langle g_t, w_t - u\rangle \leq \frac{\eta}{2}\sum_{t=1}^{T}||g_t||_2^2 + \frac{1}{2\eta}\sum_{t=1}^{T}\left(||u - w_t||^2 - ||u - w_{t+1}||^2\right)$$

By telescoping of the term in the second sum, we can cancel all the terms except the first (as ever other term will appear with a positive and then a negative sign) and dropping the final term, we are left with

$$\sum_{t=1}^{T}\langle g_t, w_t - u\rangle \leq \frac{\eta}{2}\sum_{t=1}^{T}||g_t||_2^2 + \frac{1}{2\eta}||u - w_1||^2$$

□

2 Online Mirror Descent

Motivating Question:
Can we develop algorithms with regrets that scale with other geometric measures of data (e.g. ℓ_∞,ℓ_1, etc.)?

2.1 Background on norms

Definition 4. *A function, $||\cdot||$, ($\mathbb{R}^D \to \mathbb{R}$) is said to be a **norm** if*

1. *Homogeneity: $\forall a \in \mathbb{R}$ and $x \in \mathbb{R}^d$, then $||ax|| = |a|||x||$*
2. *Triangle Inequality: $\forall x, y \in \mathbb{R}^d$, $||x + y|| \leq ||x|| + ||y||$*
3. *Point Separation: $||x|| = 0 \Rightarrow x = \vec{0}$*

Examples

1. $||x||_2 = \sqrt{x_1^2 + \cdots + x_d^2}$
2. $||x||_\infty = \max_i |x_i|$
3. $||x||_p = \left(\sum_{i=1}^{p} |x_i|^p\right)^{\frac{1}{p}}$
4. Mahalanobis Norm: A is positive definite $A = PP^T$ for invertible P

$$||x||_A = \sqrt{x^T A x} = \sqrt{x^T P P^T x} = ||P^T x||_2$$

Definition 5. *Given a norm, $||\cdot||$, define the **dual norm**, $||\cdot||_*$, as*

$$||z||_* = \sup_{x:||x||\leq 1} \langle x, z\rangle$$

$\|\cdot\|$	$\|\cdot\|_*$
$\|\cdot\|_2$	$\|\cdot\|_2$
$\|\cdot\|_1$	$\|\cdot\|_\infty$
$\|\cdot\|_p \quad p \in [1, \infty]$	$\|\cdot\|_q \quad (\frac{1}{q} + \frac{1}{p} = 1)$
$\|\cdot\|_A$	$\|\cdot\|_{A^{-1}}$

Table 1: List of Norms on the left and their paired Dual Norm on the right

1. *$||\cdot||_*$ is also a norm*
2. *By definition of dual norm,*

$$\langle x, z \rangle = ||x|| \left\langle \frac{x}{||x||}, z \right\rangle \leq ||x|| ||z||_*$$

   This generalizes Cauchy-Schwarz

Examples

2.2 General Lipshitz Property (w.r.t. arbitrary norm)

Recall: $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lip w.r.t. $||\cdot||$ if $\forall x, y\ |f(x) - f(y)| \leq L||x - y||$

Lemma 6. *Relating Lipshitzness to gradient norm*

1. *$f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then f is L-Lip w.r.t. $||\cdot|| \Leftrightarrow \forall x\ ||\nabla f(x)||_* \leq L$*
2. *$f : \mathbb{R}^d \to \mathbb{R}$ is convex, then f is L-Lip w.r.t. $||\cdot|| \Leftrightarrow \forall x \forall g \in \partial f(x)$, $||g||_* \leq L$*

Proof. 1. Left as an exercise, proof uses ideas similar to part 2

2. $(\Rightarrow)$ $\forall x \forall g \in \partial f(x)$ we have $\forall y$

$$f(y) \geq f(x) + \langle g, y - x \rangle$$

Now pick g^* s.t. $||g^*|| \leq 1$ and $\langle g, g^* \rangle = ||g||_*$ and let $y = x + g^*$, then

$$||g||_* = \langle g, g^* \rangle = \langle g, y - x \rangle \leq f(y) - f(x) \leq L||y - x|| = L||g^*|| \leq L$$

$(\Leftarrow)$ $\forall x, y$ and $g_y \in \partial f(y)$, $g_x \in \partial f(x)$

$$-L||x - y|| \leq \langle g_y, x - y \rangle \leq f(x) - f(y) \leq \langle g_x, x - y \rangle \leq L||x - y||$$

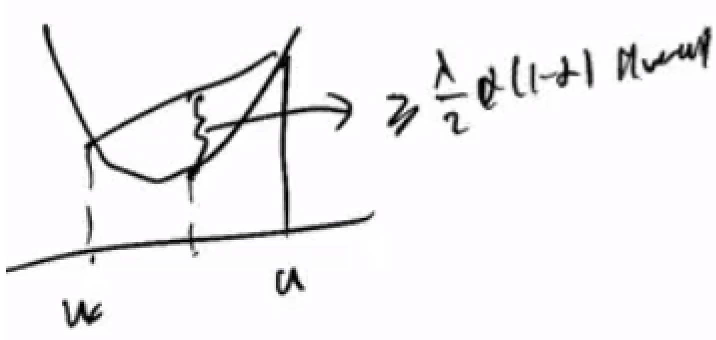
□

2.3 Strong Convexity (for arbitrary norm)

Recall: f is λ-Strongly Convex (sc) if $\forall u, w$ and $\alpha \in (0, 1)$

$$f(\alpha w + (1 - \alpha)u) \leq \alpha f(w) + (1 - \alpha)f(u) - \frac{\lambda}{2}\alpha(1 - \alpha)||w - u||^2$$

This second term is a gap that can be visualized in the following picture



Lemma 7. *If f is λ-SC, then $\forall u, w$ and $\forall g \in \partial f(u)$,*

$$f(w) - f(u) \geq \langle g, w - u \rangle + \frac{\lambda}{2}||w - u||^2$$

Proof. Let us start by using the definition of λ-SC

$$\frac{f(\alpha w + (1-\alpha)u) - f(u)}{\alpha} \leq \frac{\alpha f(w) + (1-\alpha)f(u) - \frac{\lambda}{2}\alpha(1-\alpha)||w-u||^2 - f(u)}{\alpha} = f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)||w-u||^2$$

Now we can lower bound the LHS by using the sub-gradient at point u, denoted as g

$$\frac{f(\alpha w + (1-\alpha)u) - f(u)}{\alpha} \geq \frac{\langle g, \alpha(w-u)\rangle}{\alpha} = \langle g, w - u\rangle$$

Combining these two, we get

$$\langle g, w - u\rangle \leq f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)||w-u||^2$$

Which holds for all α. Pick $\alpha = 0$

$$\langle g, w - u\rangle \leq f(w) - f(u) - \frac{\lambda}{2}||w-u||^2$$

Which is an alternative statement of the lemma □

Definition 8. *If ψ is differentiable and strongly convex, then*

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$$

is called the ***Bregman Divergence*** *induced by ψ.*

Next time: we will look at Online Mirror Descent

$$w_{t+1} = \underset{w \in \Omega}{\operatorname{argmin}} \langle \eta g_t, w\rangle + D_\psi(w, w_t)$$