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

Lecture 21: Proof of Online Mirror Descent

Lecturer: Chicheng Zhang *Scribe: Yao Zhao*

1 Guarantees of OMD

Online mirror descent provides a generalized form of the guarantee we found for OGD:

1.1 Regret for OMD

Theorem 1. *If Ψ is 1-SC with respect to some norm $\|\cdot\|$, then OMD with distance generating function Ψ and learning rate η has regret guarantee:*

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq \frac{D_\Psi(\mathbf{u}, \mathbf{w}_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|_*^2.$$

Specifically, if $D_\Psi(\mathbf{u}, \mathbf{w}_1) \leq H^2$ and $\forall t, \|\mathbf{g}_t\|_* \leq \rho$, then:

$$\eta = \frac{H}{\rho}\sqrt{\frac{1}{T}} \implies R_T(\mathbf{u}) \leq H\rho\sqrt{T}.$$

Before moving on to the proof of this theorem, we first discuss several interesting examples, which were first introduced in the last lecture.

Example 1: p-norm algorithm

Take $\Omega = \mathbb{R}^d$, $\Psi(\mathbf{w}) = \frac{1}{2(p-1)}\|\mathbf{w}\|_p^2$, $p \in (1, 2]$, which is convex. In addition, $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms, given

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Initialize $\mathbf{w}_1 = 0 \in \mathbb{R}^d$. We have regret bound as follows,

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq \frac{\|\mathbf{u}\|_p^2}{2(p-1)\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|_q^2,$$

where, $\mathbf{g}_t$ is the sub-gradient at $\mathbf{w}_t$. Furthermore, if f_t is R_q-Lipschitz ($\forall t, \|\mathbf{g}_t\|_q^2 \leq R_q$) w.r.t. $\|\cdot\|_q$, and $\|\mathbf{u}\|_p \leq B_p$. It then turns out,

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq \frac{B_p^2}{2(p-1)\eta} + \frac{\eta}{2}TR_q^2,$$

Tuning the step size η, we have

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq B_p R_q \sqrt{\frac{T}{p-1}},$$

when,

$$\eta = \frac{B_p}{R_q}\sqrt{\frac{1}{(p-1)T}}.$$

Especially, it reduces to OGD when $p = 2$.
Example 2: exponential weight algorithm
Take $\Omega = \Delta^{d-1}$, $\Psi(\mathbf{w}) = \sum_i w_i \ln w_i$. The initialization is $\mathbf{w}_1 = (\frac{1}{d}, ... \frac{1}{d})$. We have

$$D_\Psi(\mathbf{u}, \mathbf{w}_1) = \sum_i u_i \ln \frac{u_i}{\frac{1}{d}} \leq \ln d,$$

where, the last inequality holds as $u_i \in [0, 1]$.
It then turns out,

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq \frac{\ln d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_\infty^2.$$

Furthermore, if $\forall t, \|\mathbf{g}_t\|_\infty^2 \leq R_\infty$ then,

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq \frac{\ln d}{\eta} + \frac{\eta}{2} T R_\infty^2.$$

Tuning the step size η, we have

$$\forall \mathbf{u} \in \Omega : R_T(\mathbf{u}) \leq R_\infty \sqrt{T \ln d}.$$

when,

$$\eta = \frac{1}{R_\infty} \sqrt{\frac{\ln d}{T}}.$$

1.2 Proof of the regret bound

The proof follows the same procedure as OGD.

Proof. **Step 1:** linearization

$$R_T(\mathbf{u}) = \sum_{t=1}^{T} (f_t(\mathbf{w}_1) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T} < \mathbf{g}_t, \mathbf{w}_1 - \mathbf{u} > .$$

Step 2: first order optimality condition at w_{t+1},

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w} \in \Omega} < \eta \mathbf{g}_t, \mathbf{w} > + D_\Psi(\mathbf{w}, \mathbf{w}_t).$$

where, the two terms of right hand side correspond to correctiveness and conservativeness respectively. The first order optimality condition tells us that

$$< \nabla f(\mathbf{w}_{t+1}), \mathbf{u} - \mathbf{w}_{t+1} > \geq 0.$$

which means

$$\langle \nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t) + \eta \mathbf{g}_t, \mathbf{u} - \mathbf{w}_{t+1} \rangle \geq 0$$

Thus,

$$\begin{aligned}\langle \mathbf{g}_t, \mathbf{w}_{t+1} - \mathbf{u} \rangle &\leq \frac{1}{\eta} \langle \mathbf{u} - \mathbf{w}_{t+1}, \nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t) \rangle \\ &= \frac{1}{\eta} \langle D_\Psi(\mathbf{u}, \mathbf{w}_t) - D_\Psi(\mathbf{u}, \mathbf{w}_{t+1}) - D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t) \rangle.\end{aligned}$$

Step 3: bounding the instantaneous $\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{u} \rangle$

$$\begin{aligned}\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{u} \rangle &= \langle \mathbf{g}_t, \mathbf{w}_{t+1} - \mathbf{u} \rangle + \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_{t+1} \rangle \\ &= \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_*^2 + \frac{1}{\eta}(D_\Psi(\mathbf{u}, \mathbf{w}_t) - D_\Psi(\mathbf{u}, \mathbf{w}_{t+1})).\end{aligned}$$

Step 4: sum over t

$$\begin{aligned}R_T(\mathbf{u}) &\le \sum_{t=1}^{T} < \mathbf{g}_t, \mathbf{w}_t - \mathbf{u} > \\ &\le \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_*^2 + \frac{1}{\eta} \sum_{t=1}^{T} (D_\Psi(\mathbf{u}, \mathbf{w}_t) - D_\Psi(\mathbf{u}, \mathbf{w}_{t+1})) \\ &\le \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_*^2 + \frac{1}{\eta} D_\Psi(\mathbf{u}, \mathbf{w}_1).\end{aligned}$$

□

1.3 Example

A concrete example is provided here. Consider the weather prediction problem where there are d experts. Take $\Omega = \Delta^{d-1}$ and $f_t(\mathbf{w}) = \langle \mathbf{w}, l_t \rangle$. And $\forall t, i = 1, \ldots, d, l_t(t) \in [0, 1]$. How do we design OMD to minimize $R_t(\Omega)$? Here we illustrate the results with the following algorithms that we have seen before.

1. exponential weight algorithm,

$$R_T(\Omega) \le \frac{\ln d}{\eta} + \eta \sum_{t=1}^{T} \|l_t\|_\infty^2 \le O(\sqrt{T \ln d}).$$

2. online gradient decent,

$$R_T(\Omega) \le \frac{\max_{\mathbf{u}, \mathbf{v} \in \Omega} \|\mathbf{u} - \mathbf{v}\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|l_t\|_2^2.$$

Where, the first term is bounded by a constant, and the second term is bounded by $T\eta d$, which is tight at $(1, \ldots, 1)$. By tuning the step size η, we regret bound of online gradient decent is upper bounded by $O(\sqrt{Td})$.

3. p-norm algorithm,

Take $p = \frac{\ln d}{\ln d - 1}$ and $q = \ln d$.

$$R_T(\Omega) \le \frac{\max_{\mathbf{u} \in \Omega} \|\mathbf{u}\|_p^2}{2\eta(p-1)} + \frac{\eta}{2} \sum_{t=1}^{T} \|l_t\|_q^2 \le \frac{1}{2\eta}(\ln d - 1) + \frac{\eta}{2} e^2 T \le O(\sqrt{\ln d \cdot T})$$

where, $\|l_t\|_q^2 = \left(\sum_{i=1}^{d} l_t(i)^q \right)^{\frac{1}{q}} \le d^{\frac{1}{q}} \le e$

2 Some more general OCO results

2.1 Design algorithm when T is unknown

1. Doubling trick. Suppose you are given an algorithm that accepts the horizon T as parameter and has regret guarantee of $a\sqrt{T}$. Let $T_1 < T_2 < \dots$ be a fixed sequence of integers and consider the algorithm that runs with horizon T_1 until $t = T_1$, then runs with horizon T_2 until $t = T_1 + T_2$, and then restart again with horizon T_1 until $t = T_1 + T_2 + T_3$, where $T_{l+1} = 2T_l$. The resulting regret bound is $\frac{\sqrt{2}}{\sqrt{2}-1} a\sqrt{t}$.

2. Time-varying step size. By using step size $\eta_t = \frac{H}{\rho}\sqrt{\frac{1}{t}}$, we can achieve regret bound as $O(H\rho\sqrt{T})$.

2.2 Optimality of regret guarantee

Theorem 2. *Let $\Omega \in \mathbb{R}^d$ be a convex set, and $D = \sup_{\mathbf{u},\mathbf{v}\in\Omega} \|\mathbf{u} - \mathbf{v}\|_2$. For any algorithm $\mathcal{A}$, and time horizon T. Then there exists a sequence of linear functions $f_t(\mathbf{w}_t) = \langle \mathbf{g}_t, \mathbf{w} \rangle$, and $\|\mathbf{g}_t\|_2 \le L$, such that*

$$R_T(\Omega) \ge \frac{LD\sqrt{T}}{4}$$

This lower bounds shows that OGD is optimal for Ω, and assuming L-Lipschitz of the loss functions w.r.t. $\|\cdot\|_2$.

We can also show a similar result for $\Omega \in \Delta^{d-1}$, which means the exponential weight algorithm is optimal.

$$R_T(\Omega) \ge const \cdot L\sqrt{T \ln d}$$

Here, we need the assumption that T is large enough, and $\|\mathbf{g}_t\|_\infty \le L$.

Caveat: This doesn't rule out algorithms that can exploit "easy data" or "weak adversary".

2.3 Follow the regularized leader

At each time step, $t = 1, 2, ..., T$:

- choose $\mathbf{w}_t = \arg\min_{\mathbf{w}\in\Omega} \sum_{s=1}^{t-1} f_s(\mathbf{w}) + \frac{\Psi(\mathbf{w})}{\eta}$;
- receive f_t and suffer the loss $f_t(\mathbf{w_t})$;

Intuitively, $\mathbf{w}_t$ will oscillate a lot without regularization ($\eta \to \infty$, aka FTL), which induces large regret. But with this regularization, the algorithm becomes stable and amenable for analysis.

Remark 3. *Chicheng notes after lecture: sometimes people consider doing a first order approximation on f_s's. This results in an OCO algorithm that only need to access the subgradients of the loss functions. This is called Nesterov's dual averaging. Specifically:*

At each time step, $t = 1, 2, ..., T$:

- *choose $\mathbf{w}_t = \arg\min_{\mathbf{w}\in\Omega} \sum_{s=1}^{t-1} \langle g_s, w \rangle + \frac{\Psi(\mathbf{w})}{\eta}$;*
- *receive f_t and suffer the loss $f_t(\mathbf{w_t})$, receive $g_t \in \partial f_t(\mathbf{w_t})$;*

In some setting of Ω and Ψ, FTRL may coincide with OMD,

$$\mathbf{w}_t = \arg\min_{\mathbf{w}\in\Omega} \sum_{s=1}^{t-1} \langle \mathbf{g}_s, \mathbf{w} \rangle + \frac{\Psi(\mathbf{w})}{\eta} = \nabla\Psi_\Omega^* \left(-\eta \sum_{s=1}^{t-1} \mathbf{g}_s \right).$$

Next lecture will be about exploiting strong convexity in online convex optimization.