HA 2

Modern Method of Decision Making

Work has been done by: Ryabykin Aleksey, Osokin Mikhail

Adaptive Learning Rates for Online Gradient Descent

Task Description

Consider:

- a closed and convex decision space K with diameter:

$$\mathrm{Diam}(K) = \max_{x,y\in K} \|x-y\|_2 \leq D.$$

- a loss class $\mathcal{F}$ of convex and differentiable $f : K \to \mathbb{R}$

The goal of this home assignment is to study, in the context of the Online Convex Optimization problem, the perfomance of the Adaptive Online Gradient Descent algorithm defined as follows:

Initialize:
  $x_1 \in K$
for $t \geq 1$ **do**
  → Play x_t;
  → Receive loss f_t;
  → Incur loss $f_t(x_t)$;
  → Update:

$$x_{t+1} = \pi_K(x_t - \gamma_t \nabla_t), \quad \nabla_t = \nabla f_t(x_t)$$

  where:

$$\gamma_t = \frac{D}{\sqrt{\sum_{s=1}^{t} \|\nabla_s\|_2^2}}$$

end for

Note

In the sequel, we denote:

$$R_T = \sum_{t=1}^{T} f_t(x_t) - \inf_{x\in K} \sum_{t=1}^{T} f_t(x)$$

Solution

1. **Show that, for any value of the learning rates γ_t, we have:**

$$R_T \leq \frac{D^2}{2\gamma_T} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|\nabla_t\|_2^2.$$

Proof:

Lemma

Fix $x_1 \in K$ as well as positive numbers $(\gamma_t)_{t\geq 1}$. For any sequence $(g_t)_{t\geq 1}$ of vectors in $\mathbb{R}^d$, define

$$x_{t+1} = \pi_k(x_t - \gamma_t g_t), \quad t \geq 1.$$

Then, $\forall T \geq 1,\ \forall x \in K$:

$$\sum_{t=1}^{T} \langle g_t, x_t - x\rangle \leq \frac{1}{2}\sum_{t=1}^{T}\left\{\left(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right)\|x_t - x\|_2^2 + \gamma_t\|g_t\|_2^2\right\}$$

$\forall x \in K,\ \forall t \geq 1$, we deduce by convexity of f_t and definition of ∇_t, that:

$$f_t(x_t) - f_t(x) \leq \langle \nabla_t, x_t - x\rangle.$$

As a result,

$$\sum_{t=1}^{T} f_t(x_t) - \inf_{x\in K}\sum_{t=1}^{T} f_t(x) = \sup_{x\in K}\sum_{t=1}^{T}\left(f_t(x_t) - f_t(x)\right) \leq \sup_{x\in K}\sum_{t=1}^{T}\langle \nabla_t, x_t - x\rangle.$$

According to the lemma above, and the definition of D, we deduce that:

$$\sum_{t=1}^{T} f_t(x_t) - \inf_{x\in K}\sum_{t=1}^{T} f_t(x) \leq \sup_{x\in K}\frac{1}{2}\sum_{t=1}^{T}\left\{\left(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right)\underbrace{\|x_t - x\|_2^2}_{\leq D^2} + \gamma_t\|\nabla_t\|_2^2\right\} \leq$$

$$\leq \frac{D^2}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right) + \frac{1}{2}\sum_{t=1}^{T}\gamma_t\|\nabla_t\|_2^2$$

$$= \frac{D^2}{2\gamma_T} + \frac{1}{2}\sum_{t=1}^{T}\gamma_t\|\nabla_t\|_2^2$$

□

2. **Show that if $\phi : (0, +\infty) \to \mathbb{R}$ is a non-increasing function and $(u_t)_{t\geq 1}$ are positive numbers, then $\forall T \geq 1$:**

$$\sum_{t=1}^{T} u_t \phi\left(\sum_{s=1}^{T} u_s\right) \leq \int_0^{\sum_{s=1}^{T} u_s} \phi(\omega)\, d\omega$$

Proof: Let $a, b \in \mathbb{R}^{+}$. Let us prove that $b \cdot \phi(a+b) \leq \int\limits_{a}^{a+b} \phi(\omega)d\omega$. The integral on the right side is equal to the area under ϕ on the segment $[a, a+b]$. ϕ is a non-increasing function, which means that $\phi(a+b)$ is the minimal value of ϕ on the entire segment. If ϕ was equal to $\phi(a+b)$ on the whole segment, then the area would be $b \cdot \phi(a+b)$. Since the value of ϕ can only go higher from $\phi(a+b)$ on the segment, the area under ϕ there is bounded from below by $b \cdot \phi(a+b)$, which concludes the proof. Now, if we set a to 0 and b to u_1, we get $u_1\phi(u_1) \leq \int\limits_{0}^{u_1} \phi(\omega)d\omega$, which is the induction base.

Also, if $\sum\limits_{t=1}^{T-1} u_t\phi(\sum\limits_{s=1}^{t} u_s) \leq \int\limits_{0}^{\sum\limits_{s=1}^{t-1} u_s} \phi(\omega)d\omega$, then

$$\sum_{t=1}^{T-1} u_t\phi\left(\sum_{s=1}^{t} u_s\right) + u_T \cdot \phi\left(\sum_{s=1}^{T} u_s\right) \leq \int\limits_{0}^{\sum\limits_{s=1}^{T-1} u_s} \phi(\omega)d\omega + \int\limits_{\sum\limits_{s=1}^{T-1} u_s}^{\sum\limits_{s=1}^{T} u_s} \phi(\omega)d\omega$$

and

$$\sum_{s=1}^{T} u_t\phi\left(\sum_{s=1}^{t} u_s\right) \leq \int\limits_{0}^{\sum\limits_{s=1}^{T} u_s} \phi(\omega)d\omega, \quad \text{since } u_1 \cdot \phi\left(\sum_{s=1}^{t} u_s\right) \leq \int\limits_{\sum\limits_{s=1}^{T-1} u_s}^{\sum\limits_{s=1}^{T} u_s} \phi(\omega)d\omega$$

which is the induction step. ☐

3. Suppose $\gamma_t = \dfrac{D}{\sqrt{\sum\limits_{s=1}^{t} \|\nabla_s\|_2^2}}$. Combining 1. and 2. (for a well chosen value of u_t and ϕ) show that:

$$R_T \leq \frac{3D}{2}\sqrt{\sum_{t=1}^{T} \|\nabla_t\|_2^2}$$

Proof: Let's substitute γ_t to the result of first step.

$$R_T \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T} \|\nabla_t\|_2^2} + \frac{D}{2}\sum_{t=1}^{T} \frac{\|\nabla_t\|_2^2}{\sqrt{\sum\limits_{s=1}^{t} \|\nabla_s\|_2^2}}.$$

Let $\phi(n) = \frac{1}{\sqrt{n}}$, $u_t = ||\nabla_t||_2^2$, then $\sum\limits_{t=1}^{T} \frac{u_t}{\sqrt{\sum\limits_{s=1}^{T} u_s}} \leq \int\limits_{0}^{\sum_{s=1}^{T} u_s} \frac{1}{\sqrt{n}} dn = 2\sqrt{\sum\limits_{s=1}^{T} u_s} = 2\sqrt{\sum\limits_{t=1}^{T} ||\nabla_t||_2^2}$. This means that:

$$R_T \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T} ||\nabla_t||_2^2} + \frac{D}{2}\sum_{t=1}^{T} \frac{||\nabla_t||_2^2}{\sqrt{\sum\limits_{s=1}^{t} ||\nabla_s||_2^2}} \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T} ||\nabla_t||_2^2} + \frac{D}{2} \cdot 2\sqrt{\sum_{t=1}^{T} ||\nabla_t||_2^2} =$$
$$= \frac{3D}{2}\sqrt{\sum_{t=1}^{T} ||\nabla_t||_2^2}.$$

□

4. Explain why this is always a better performance guarantee than the one we provided for the OGD algorithm studied in Lecture 4:

Theorem

Suppose that

- (Bounded diameter): $\mathrm{Diam}(K) \leq D < +\infty$;
- (Bounded subgradients):

$$\forall x \in K,\ \forall f \in \mathcal{F},\ \forall \nabla \in \partial f(x):\ ||\nabla||_2 \leq G < +\infty.$$

Then $\forall x_1 \in K$, the OGD algorithm with step size $\gamma_t = \frac{D}{G\sqrt{t}}$, $\forall t \geq 1$, satisfies:

$$R_T \leq \frac{3}{2} GD\sqrt{T}.$$

The inequality above can be written in the following way:

$$R_{T_{Lec4}} \leq \frac{3}{2} GD\sqrt{T} = \frac{3D}{2}\sqrt{TG^2} = \frac{3D}{2}\sqrt{\sum_{t=1}^{T} G^2}$$

Since $\forall \nabla \in \partial f(x)$ $||\nabla||_2 \leq G$ we can obtain, that:

$$R_T \leq \frac{3D}{2}\sqrt{\sum_{t=1}^{T} ||\nabla_t||_2^2} \leq \frac{3D}{2}\sqrt{\sum_{t=1}^{T} G^2} \leq R_{T_{Lec4}}.$$

5. Recall that a function $f : K \to \mathbb{R}$ is called β-smooth if it is differentiable and such that $||\nabla f(x) - \nabla f(y)||_2 \leq \beta ||x - y||_2$, $\forall x, y \in K$..

Show that if $f : K \to \mathbb{R}$ is β-smooth and achieves a minimum at $x^* \in K$, then $\forall x \in K$:

$$||\nabla f(x)||_2^2 \leq 2\beta(f(x) - f(x^*)).$$

Proof: Since smoothness and the optimality of x^* ($f : K \to \mathbb{R}$ is β-smooth and achieves minimum at $x^* \in K$), we have:

$$f(x^*) \le f\left(x - \frac{1}{\beta}\nabla f(x)\right) \le f(x) - \frac{1}{\beta}\|\nabla f(x)\|_2^2 + \frac{1}{2\beta}\|\nabla f(x)\|_2^2 \le$$

$$f(x) - \frac{1}{2\beta}\|\nabla f(x)\|_2^2 \Rightarrow f(x^*) \le f(x) - \frac{1}{2\beta}\|\nabla f(x)\|_2^2$$

Multiplying both sides by 2β we obtain:

$$\|\nabla f(x)\|_2^2 \le 2\beta \cdot (f(x) - f(x^*)) .$$

□

6. Suppose now that the losses that the losses $f \in \mathcal{F}$ are also β-smooth and positive. Combine 3. and 5. to show that:

$$R_T \le \sqrt{\frac{9D^2\beta}{2}\sum_{t=1}^{T} f_t(x_t)}$$

Proof:

$$\frac{3D}{2}\sqrt{\sum_{t=1}^{T}\|\nabla_t\|_2^2} \le \frac{3D}{2}\sqrt{2\beta\sum_{t=1}^{T}(f_t(x_t) - f_t(x^*))} \le$$

Since the loss is positive:

$$\le \frac{3D}{2}\sqrt{2\beta\sum_{t=1}^{T} f_t(x_t)} = \sqrt{\frac{9D^2\beta}{2}\sum_{t=1}^{T} f_t(x_t)}.$$

□

7. Conclude from 6. that if the losses are β-smooth and positive, then:

$$R_T \le \frac{9D^2\beta}{2} + 2\sqrt{\frac{9D^2\beta}{2}\inf_{x\in K}\sum_{t=1}^{T} f_t(x_t)}$$

Proof: As we denote:

$$R_T = \sum_{t=1}^{T} f_t(x_t) - \inf_{x\in K}\sum_{t=1}^{T} f_t(x)$$

We can substitute it in the result of 6.:

$$R_T \leq \sqrt{\frac{9D^2\beta}{2}\sum_{t=1}^{T} f_t(x_t)} =$$

$$= \sqrt{\frac{9D^2\beta}{2}\left(R_T + \inf_{x\in K}\sum_{t=1}^{T} f_t(x)\right)} \Longrightarrow$$

$$\Rightarrow R_T^2 - \frac{9D^2\beta}{2}\cdot R_T - \frac{9D^2\beta}{2}\cdot \inf_{x\in K}\sum_{t=1}^{T} f_t(x) \leq 0$$

$$\mathcal{D} = \left(\frac{81D^4\beta^2}{4}\right) + 4\frac{9D^2\beta}{2}\cdot \inf_{x\in K}\sum_{t=1}^{T} f_t(x)$$

Hence

$$R_T^+ = \frac{\frac{9D^2\beta}{2} + \sqrt{\left(\frac{81D^4\beta^2}{4}\right) + 4\frac{9D^2\beta}{2}\cdot \inf_{x\in K}\sum_{t=1}^{T} f_t(x)}}{2}$$

So

$$R_T \leq \frac{9D^2\beta}{4} + \frac{9D^2\beta}{4} + 2\sqrt{\frac{9D^2\beta}{2}\inf_{x\in K}\sum_{t=1}^{T} f_t(x)} = \frac{9D^2\beta}{2} + 2\sqrt{\frac{9D^2\beta}{2}\inf_{x\in K}\sum_{t=1}^{T} f_t(x)}$$

□