Advanced Topics in Machine Learning and Algorithmic Game Theory

Fall semester, 2011/12

Lecture 6: MAB Online Convex Optimization:

Lecturer: Yishay Mansour *Scribe: Orr Tamir*

6.1 Lecture Overview

In this lecture we turn our attention to the online convex optimization in the *multi-armed bandit* (MAB) model. In this model, there is a set N of actions from which the player has to choose in step $t \in T$. After choosing the action, the player can only see the loss of her action, not the losses of the other possible actions. We will consider convex problem $c : \Re^d \to \Re$. We will use gradient decent to solve that problem with $O(T^{5/6})$ regret.

Later on we will consider the difference between Adaptive and Oblivious Opponents by showing an example of Adaptive opponent for EXP3 that gets $\Omega(T^{3/4})$ regret instead of $O(\sqrt{T})$ regret which we proved in Lecture 4 for Oblivious opponent.

6.2 Online Convex Optimization: MAB

The idea here is to use gradient-descent based algorithm. For convex problem $c : \Re^d \to \Re$, the gradient-decent method calculates:

$$x_{t+1} = x_t - \eta \bigtriangledown c(x_t)$$

For stochastic problem:

$$\begin{aligned} c_t(x) &= c(x) + \varepsilon_t(x) \qquad \text{where} \quad \mathbb{E}[\varepsilon_t(x)] = 0 \\ x_{t+1} &= x_t - \eta \nabla c_t(x) \end{aligned}$$

The important thing is:

$$\mathbb{E}[\bigtriangledown c_t(x)] = \bigtriangledown \mathbb{E}[c_t(x)] = \bigtriangledown c(x)$$

In the MAB world we don't have $\forall x$ the value of $c_t(x)$ we have $c_t(x_t)$ for our chosen action x_t. If we don't have a function how we will calculate the derivative?

6.2.1 Estimating gradient with one sample

For $d = 1$ (one dimension)

$$\begin{aligned} f'(x) &\approx \frac{f(x+\delta) - f(x-\delta)}{2\delta} \\ &= \frac{1}{2} \sum_{\nu \in \{1,-1\}} \frac{\nu f(x+\nu\delta)}{\delta} \\ &= \frac{1}{2} \mathbb{E}_\nu [\frac{\nu f(x+\nu\delta)}{\delta}] \end{aligned}$$

With one sample we got a stochastic approximation to the estimation of the derivative. In higher dimensions $\triangledown f(x) = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d})$. For $u = \pm e_i$(standard base vectors):

$$\triangledown f(x) \approx \mathbb{E}[d \frac{f(x+\delta u)}{\delta} u]$$

The gradient isn't base depended , so we can choose a random $||u|| = 1$ and the previous identity holds. We will show that this estimator is the gradient of $\hat{f}(x)$ (even when the gradient of f is not define).

$$\begin{aligned} \hat{f}(x) &= \mathbb{E}_{\nu \in \mathcal{B}}[\frac{d}{\delta} f(x+\delta\nu)\nu], \quad \text{where } \mathcal{B} && = \{x \mid ||x|| \leq 1\} \\ \triangledown \hat{f}(x) &= \mathbb{E}_{\nu \in \mathcal{S}}[\frac{d}{\delta} f(x+\delta\nu)], \quad \text{where } \mathcal{S} && = \{x \mid ||x|| = 1\} \end{aligned}$$

We can view $\hat{f}$ as a way to smooth f such that it is also continuous and differential.

Lemma 6.2.1 $\forall \delta > 0, \mathbb{E}_{\nu \in \mathcal{S}}[f(x+\delta\nu)] = \frac{\delta}{d} \triangledown f(x)$

For dimension $d = 1$ we have $\frac{\partial}{\partial x} \int_{-\delta}^{\delta} f(x+\nu) d\nu = f(x+\delta) - f(x-\delta), d = 1$

For general dimension d,We use Stoke theorem that states:

$$\triangledown \int_{\delta \mathcal{B}} f(x+\nu) d\nu = \int_{\delta \mathcal{S}} f(x+u) \frac{u}{\|u\|} du \tag{6.1}$$

$$\hat{f}(x) = \mathbb{E}[f(x+\delta\nu)] = \frac{\int_{\delta \mathcal{B}} f(x+\nu) d\nu}{Vol_{d-1}(\delta \mathcal{S})} \tag{6.2}$$

$$\mathbb{E}[f(x+\delta\nu)\nu] = \int_{\delta \mathcal{S}} f(x+u) \frac{u}{\|u\|} du \tag{6.3}$$

$$\frac{Vol_d(\delta \mathcal{B})}{Vol_{d-1}(\delta \mathcal{S})} = \frac{\delta}{d} \tag{6.4}$$

We have the following identities,
$\mathbb{E}[f(x+\delta\nu)] =^{6.3} \frac{\int_{\delta \mathcal{S}} f(x+u) \frac{u}{\|u\|} du}{Vol_{d-1}(\delta \mathcal{S}} =^{6.1} \frac{\triangledown \int_{\delta \mathcal{B}} f(x+\nu) d\nu}{Vol_{d-1}(\delta \mathcal{S}} = \frac{Vol_d(\delta \mathcal{B})}{Vol_{d-1}(\delta \mathcal{S})} \triangledown \hat{f}(x) =^{6.4} \frac{\delta}{d} \triangledown \hat{f}(x)$

6.3 Bandit Gradient Decent

Unknown convex function:$c_t : S \to [-M, M]$

The algorithm $BGD(\alpha, \delta, \eta)$
At each period t:

1. $x_t = y_t + \delta u_t$ where $u_t = randUnitVector()$
2. $play\ x_t$
3. observe $c_t(x_t)$
4. $y_{t+1} = \Pi_{(1-\alpha)S}(y_t - \eta c_t(x_t) u_t)$ where $\Pi_A(z)$ is a projection of z to a set A

Assumptions:

- c_t has a bounded gradient:
  $max_{x \in S} \| \bigtriangledown c_t(x_t) \| \leq G$
- S is a convex set s.t:
  $\exists r \exists R, r\mathcal{B} \subseteq S \subseteq R\mathcal{B}$, where $\mathcal{B}$ is the unit ball.

Lemma 6.3.1 *The optimum in $(1-\alpha)S$ is close to the optimum in S :*
$min_{x \in (1-\alpha)S} \sum_{t=1}^{T} c_t(x) \leq 2\alpha MT + min_{x \in S} \sum_{t=1}^{T} c_t(x)$

Proof. $\forall x \in S$, we have $(1-\alpha)x \in (1-\alpha)S$
From the fact the c_t is convex we get:
$c_t((1-\alpha)x) = c_t((1-\alpha)x + \alpha * 0) \leq (1-\alpha)c_t(x) + \alpha c(0) \leq c_t(x) + 2\alpha M$
Summing on the time steps we have:
$\sum_{t=1}^{T} c_t((1-\alpha)x) \leq 2\alpha MT + \sum_{t=1}^{T} c_t(x)$
This is true for any $x \in S$, so it holds for $x^\star = argmin_{x \in S} \sum_t c_t(x)$ □

Lemma 6.3.2 $\forall x \in (1-\alpha)S$ *the ball with radius αr where x is its center, is a subset of S.*

Proof. Mnikowsky sum of two sets is: $A + B = \{a + b | a \in A, b \in B\}$. We have

$$(1-\alpha)S + \alpha r\mathcal{B} \subseteq (1 - \alpha S) + \alpha S = S$$

where the equality golds since S is a convex set. □

Lemma 6.3.3 $\forall x \in (1-\alpha)S, y \in S$
$|c_t(x) - c_t(y)| \leq \frac{2M}{\alpha r} |x - y|$.

Proof. Define $\triangle$,s.t $y = x + \triangle$
If $| \triangle | > \alpha r$ we finished since $|c_t(x)| \leq M$. Otherwise, let $z = x + \alpha r \frac{\triangle}{\|\triangle\|}$. From previous Lemmas $z \in S$.We have

$$y = \frac{\| \triangle \|}{\alpha r} z + (1 - \frac{| \triangle |}{\alpha r})x$$

Since c_t is convex,

$$c_t(y) \leq c_t(x) + \frac{c_t(z) - c_t(x)}{\alpha r} | \triangle | \leq= c_t(x) + \frac{2M}{\alpha r}|x - y|$$

□

We will build now the proof of the BGD.

Theorem 6.1 *(correctness) $\forall x_t$ (from the algorithm) $x_t \in S$*

Proof. We have $y_t \in (1 - \alpha)S$ from Lemma 6.3.2 $x_t \in S$ for $\frac{\delta}{r} \leq \alpha \leq 1$. □

Theorem 6.2 *The regret of BGD is $O(T^{\frac{5}{6}} M \sqrt{\frac{dR}{r}})$.*

Proof. The proof will be done in two steps. we first show the regret of the y_t's w.r.t. the $\hat{c}_t$ over the set $(1 - \alpha)S$.

(step A) Regret bound for y_t, with functions $\hat{c}_t()$ and over the set $(1 - \alpha)S$. We will examine the run of the algorithm for y_t and consider a run of gradient decent for

$$\hat{c}_t(x) = \mathbb{E}_{\nu \in \mathcal{B}}[c_t(x + \delta \nu)].$$

Define: $g_t = \frac{d}{\delta} c_t(y_t + \delta u_t)u_t$. From Lemma 6.2.1 $\bigtriangledown \hat{c}_t(y_t) = \mathbb{E}[g_t|y_t]$.The Update rule:

$$y_{t+1} = \Pi_{(1-\alpha)S}(y_t - \eta^* g_t) = \Pi_{(1-\alpha)S}(y_t - \eta^* \frac{d}{\delta} c_t(y_t + \delta u_t)u_t)$$

For $\eta^* = \eta \frac{\delta}{d}$ we will get our update rule.We will bound the gradient:

$$|g_t| = |\frac{d}{\delta} c_t(y_t + \delta u_t)u_t| \leq \frac{d}{\delta} M \triangleq G$$

From Stochastic Gradient Decent result we will get:

$$\mathbb{E}[\sum_{t=1}^{T} \hat{c}_t(y_t)] - min_{y \in (1-\alpha)S} \sum_{t=1}^{T} \hat{c}_t(y) \leq R \frac{dM}{\delta} \sqrt{T}$$

We show that for $L = \frac{2M}{\alpha r}$ it by Lemma 6.3.3 . $|c_t(x) - c_t(y)| \leq L|x - y|$.For $x \in (1 - \alpha)S$, we have $|\hat{c}_t(x) - c_t(x)| \leq \delta L$. In addition, from Lemma 6.3.3 we have,

$$|\hat{c}_t(y) - c_t(x)| \leq |\hat{c}_t(y_t) - c_t(y_t)| + |c_t(y_t) - c_t(x_t)| \leq 2\delta L$$

which implies that $c_t(x_t) - 2\delta L \leq \hat{c}_t(y_t)$ and $\hat{c}_t(y) \leq c_t(x) + \delta L$. In Step A we showed:

$$\mathbb{E}[\sum_{t=1}^{T} \hat{c}_t(y_t)] - min_{y \in (1-\alpha)S} \sum_{t=1}^{T} \hat{c}_t(y) \leq R\frac{dM}{\delta}\sqrt{T}$$

Using the bounds on $\hat{c}_t(y_t)$ we have:

$$\mathbb{E}[\sum_{t=1}^{T} c_t(x_t) - 2\delta L] - min_{y \in (1-\alpha)S} \sum_{t=1}^{T} \hat{c}_t(y) \leq R\frac{dM}{\delta}\sqrt{T}$$

$$\mathbb{E}[\sum_{t=1}^{T} c_t(x_t) - 2\delta L] - min_{x \in S} \sum_{t=1}^{T} c_t(x) \leq R\frac{dM}{\delta}\sqrt{T} + 3\delta LT + 2\alpha MT = R\frac{dM}{\delta}\sqrt{T} + 3\delta\frac{2M}{\alpha r}T + 2\alpha MT$$

The regret is bounded by,

$$R\frac{dM}{\delta}\sqrt{T} + 6\delta\frac{M}{\alpha r}T + 2\alpha MT$$

We need to set δ and α, and for this we solve:

$$\min_{\delta,\alpha}(\frac{a}{\delta} + \frac{\delta}{\alpha}b + \alpha c)$$

The optimal parameters values $\delta = \sqrt[3]{\frac{a^2}{bc}}$,$\alpha = \sqrt[3]{\frac{ab}{c^2}}$ which implies a bound of $3\sqrt[3]{abc} = O(T^{\frac{5}{6}}M\sqrt{\frac{dR}{r}})$ The resulting regret is parameters are $\eta = \frac{R}{M\sqrt{T}} = O(\frac{1}{\sqrt{T}})$, $\delta = \sqrt[3]{\frac{rR^2d^2}{12T}} = O(\frac{1}{\sqrt[3]{T}})$, $\alpha = \sqrt[3]{\frac{Rd}{2r\sqrt{T}}} = O(\frac{1}{\sqrt[6]{T}})$ □

6.4 Adaptive vs Oblivious Opponent

If we can simulate the algorithm they are equal. Therefore for any deterministic algorithm there is no difference. We shall look on example for adaptive adversary for EXP3. In out case $k = 2$(only two actions). Let p be the probability of action 1 in EXP3.

$$g(t) = \begin{cases} (1,0) & \text{if } p < \alpha, \\ (0,1) & \text{if } p \geq \alpha \end{cases}$$

It's easy to see that the probability p will stay near α that's since
$p_t \leq \alpha \Rightarrow p_{t+1} \geq p_t$
$p_t \geq \alpha \Rightarrow p_{t+1} \leq p_t$
Let $\eta = \frac{1}{\sqrt{T}}$(not critical) s.t EXP3 regret is $O(\sqrt{T})$. Let $\alpha = 3\eta$
. With high probability $p_t \in [2\eta, 4\eta]$

Description of the run: We have periods, when $p_t > \alpha$,p_t will get down almost in every step until $p_t < \alpha$ then we will have a big jump when choosing action 1 (but it will take a while). In every period EXP3 gain is $1 + \frac{1}{\alpha}$ The question is how much time a period last? and how much each action gains, since this is what will determine the regret.

Let $G(q)$ be a geometrically distributed random variable with a probability q.

- The big jump: r.v $BJ\ G(\alpha), \mathbb{E}[BJ] = \frac{1}{\alpha}, Var(BJ) = \frac{1}{\alpha^2}$
- Slow get down: r.v $GD\ G(1-\alpha), \mathbb{E}[GD] = \frac{1}{\alpha}, Var(GD) = \frac{1}{\alpha}$

We will examine $O(\alpha T)$ periods. The gain of action 2: sum of $\frac{\alpha T}{2}$ r.v $G_t(\alpha)$, geometric random variables with probability α. The gain of action 2 is $\sum G_t$. For this we have, $\mathbb{E}[\sum G_t] = \frac{T}{2}$, and $Var(\sum G_t) = \frac{T}{\alpha}$. This implies that with constant probability we have a $\sqrt{\frac{T}{\alpha}}$ difference from the expectation. This will dominate the regret term, and give a regret of $T^{3/4}$.

Bibliography

[1] A.D. Falxman, A. Tauman Kalai, and H. Brendan McMahan, *Online convex optimization in the bandit setting:gradient descent without a gradient*, SODA J. Comput. Vol 32, No. 1, 2005, pp. 385-395.