IEOR 8100-001: Learning and Optimization for Sequential Decision Making 03/19/16

Lecture 15: Stochastic MABs with Global Concave Objective

Instructor: Shipra Agrawal Scribed by: Rishabh Dudeja

In the previous lecture we saw that we can reduce the bandits with constraints problem to bandits with a concave objective. In this lecture, we study an algorithm to solve this problem.

1 Problem Definition

As before consider the stochastic multi-armed bandit problem with N arms. Suppose we are given a concave $L-$Lipchitz function $f : [0,1]^d \to \mathbb{R}$. At time t, we are required to select an arm to pull, I_t. On pulling the arm we observe reward $v_t \in [0,1]^d$ generated from the distribution corresponding to the selected arm with mean V_{I_t}. That is $\mathbb{E}[v_t | I_t = i] = V_i$. The goal is to maximize $f(\frac{\sum_{t=1}^T v_t}{T})$ Now to define our notion of regret we need to select an appropriate bench-mark. We compare our algorithm against OPT_f, the reward achieved by the best static policy i.e. a fixed distribution over arms:

$$OPT_f = \max_{\boldsymbol{p}:\sum_{i=1}^N p_i = 1} f\left(\sum_{i=1}^N p_i V_i\right) \tag{1}$$

A nice property of this benchmark is that it can be shown using similar arguments as discussed in the last lecture that OPT_f is greater than the reward achieved by any dynamic policy. We define our regret $R(T)$ as:

$$R(T) = T \cdot OPT_f - Tf\left(\frac{\sum_{t=1}^T v_t}{T}\right) \tag{2}$$

Note that we have assumed that the function is $L - Lipchitz$ i.e. it cannot change very rapidly. A function is $L - Lipchitz$ w.r.t a norm $\|\cdot\|$ if $|f(x) - f(y)| \leq L\|x - y\|$. Note that in this lecture we will take $\|\cdot\|$ to be the L_2 norm.

How do we achieve low-regret for this problem? First, notice since f is a general concave L-Lipchitz function, $f\left(\frac{\sum_{t=1}^T v_t}{T}\right) \neq \frac{\sum_{t=1}^T f(v_t)}{T}$. This means that greedily maximizing $f(v_t)$ gives us no guarantee about how well we are doing w.r.t. our objective. Hence, at the end of each round we don't have immediate feedback about how well we are optimizing our objective. The trick to solving this issue will be using linearization to decompose f and obtain approximate feedback at the end of each round. To do so, first we will introduce Fenchel Duality. Then, we will solve the problem in the full information before taking decision setting. And finally we will extend our results to the bandit setting.

2 Fenchel Duality

Definition 1 (Fenchel Dual)**.** *Let $f : [0,1]^d \to \mathbb{R}$ be a concave L-Lipchitz function. The fenchel dual f^* is defined as $f^*(\theta) = \max_y \theta^T y + f(y)$*

One intuitive property of the Fenchel dual is its relation to the tangent of the function.

Fact 2. *The line $f^*(\theta) - \theta^T x$ is a tangent line to f at its point of contact for all θ.*

This is because of the following argument. Let y_θ maximize $\theta^T y + f(y)$. Because f is concave, it must be a stationary point. That is, $\theta + f'(y_\theta) = 0$. It is easy to see $f^*(\theta) - \theta^T x$ is tangent to $f(x)$ at point $x = y_\theta$ by

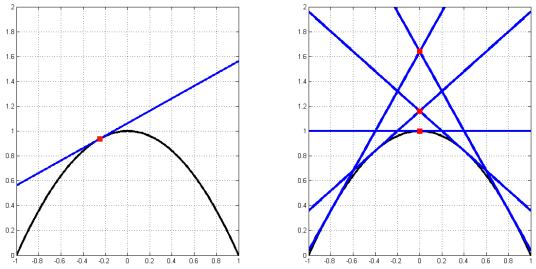
checking the line and the function have the same value as well as gradient at y_θ. Another important fact about the fenchel dual is:

Fact 3. $f(x) = \min_{\|\theta\| \leq L} f^*(\theta) - \theta^T x$

The proof of this fact is also simple: Because f is concave all tangents $f^*(\theta) - \theta^T x$ lie above $f(x)$. And hence at a point x, the lowest value among all tangents is attained by the line that is tangent precisely at x itself. And the minimizing value is $f(x)$. Finally since f is L-Lipchitz, its first derivative is bounded by L and hence we don't lose anything by maximizing over tangents with norm of the slope atmost L.

A consequence of the above fact is that at any given point x, there exists a θ_x such that, $f(x) = f^*(\theta_x) - \theta_x^T(x)$. This is ofcourse the approximation of the function with a tangent at point x.

Figure 1: Left: The line $f^*(\theta) - \theta^T x$ is a tangent line to $f(x)$. Right: $f(x) = \min_{\|\theta\| \leq L} f^*(\theta) - \theta^T x$



3 An Algorithm for the Full Information Before Taking Decision Setting

Consider the full information before taking decision setting. At time t, we observe the rewards of all arms $v_{i,t}$ before selecting an arm. We are then required to pull an arm I_t. We will first provide an algorithm for the full information setting.

Let us apply the properties of Fenchel Dual to our objective function for the problem. Hence given a sequence of rewards $v_t, \exists \bar{\theta}$ s.t. $f\left(\frac{\sum_{t=1}^T v_t}{T}\right) = f^*(\bar{\theta}) - \bar{\theta}^T \frac{\sum_{t=1}^T v_t}{T}$. Note that though we suppress this in the notation, $\bar{\theta}$ depends on the sequence of rewards we observe and is hence available only in hindsight. Suppose we knew $\bar{\theta}$. The benefit we have obtained from doing so is that maximizing the objective is same as minimizing $\bar{\theta}^T \frac{\sum_{t=1}^T v_t}{T}$. And hence, we have feedback at each step and we can hope to achieve low-regret by algorithms which greedily minimize $\bar{\theta}^T v_t$. However the downside is, that in reality $\bar{\theta}$ is not known to us. The way we get around this is by designing an algorithm which at each step t, first creates a good prediction of $\bar{\theta}$. It then selects an arm using $I_t = \arg\max_{i \in 1 \ldots N} f^*(\theta_t) - \theta_t^T v_{i,t}$. In the next section we discuss what is a good way to predict $\bar{\theta}$.

3.1 The Prediction Step

We can hope that our algorithm achieves low regret if the objective we are maximizing at each step is a good approximation to our global objective. Infact we require:

$$\sum_{t=1}^T f^*(\theta_t) - \theta_t^T v_t \leq \sum_{t=1}^T f^*(\bar{\theta}) - \bar{\theta}^T v_t + O(\sqrt{T}) \tag{3}$$

Note by the definition of $\bar{\theta}$, $\sum_{t=1}^T f^*(\bar{\theta}) - \bar{\theta}^T v_t = f(\frac{\sum_{t=1}^T v_t}{T})$. And hence the display above says that the cumulative sum of local objectives should be no more than $O(\sqrt{T})$ off from the global objective. Our requirement is very

similar to the Online Convex Optimization (OCO) problem. The theorem below defines the OCO problem and gives the regret guarantee for Online Gradient Descent (OGD).

Theorem 4 (Online Convex Optimization)**.** *Consider the following setup: At time t, we are required to choose $x_t \in A$, A, convex. We then observe a convex function $h_t(.)$ and get reward $h_t(x_t)$. Define the regret $R(T) = \sum_{t=1}^{T} h(x_t) - \min_{x \in A} \sum_{t=1}^{T} h_t(x)$. Suppose $\|\nabla h_t\| \leq G, \|x_t\| \leq D$. Then, Gradient Descent achieves a regret of atmost $O(DG\sqrt{T})$.*

Now suppose we apply Online Gradient Descent to our prediction problem. Here $x_t = \theta_t, h_t(x_t) = f^*(x_t) - x_t^T v_t$. And from the regret bound we get:

$$\sum_{t=1}^{T} h_t(x_t) = \sum_{t=1}^{T} f^*(\theta_t) - \theta_t^T v_t \tag{4}$$

$$\leq \min_{\theta} \sum_{t=1}^{T} f^*(\theta) - \theta^T v_t + O(DG\sqrt{T}) \tag{5}$$

$$\leq \sum_{t=1}^{T} f^*(\bar{\theta}) - \bar{\theta}^T v_t + O(DG\sqrt{T}) \tag{6}$$

Which is as we had required. Hence the algorithm for full information before decision setting is shown below:

for $t = 1 : T$ **do**
  Observe rewards $v_{i,t}$;
  $I_t = \arg\max_{i \in 1 \ldots N} f^*(\theta_t) - \theta_t^T v_{i,t}$;
  Play arm I_t ;
  Observe $h_t(x) = f^*(x) - x^T v_t$. Predict θ_{t+1} using OGD.
end

Algorithm 1: Online Stochastic Convex Optimization: Full Information Before Decision Setting

3.2 Regret Analysis

We first prove the following lemma:

Lemma 5. $T \cdot OPT_f \leq \mathbb{E}[\sum_{t=1}^{T} f^*(\theta_t) - \theta_t^T v_t]$

Proof. Let v_t^* be the sequence of rewards received by the OPT (best static strategy) which plays arm i with probability p_i. Let $\bar{V} = \sum_{i=1}^{N} V_i p_i$. Since v_t is chosen to maximize $f^*(\theta_t) - \theta_t^T v_t$:

$$f^*(\theta_t) - \theta_t^T v_t^* \leq f^*(\theta_t) - \theta_t^T v_t$$

Taking conditional expectation given θ_t and summing over t gives us:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[f^*(\theta_t) - \theta_t^T v_t | \theta_t] &\geq \sum_{t=1}^{T} \mathbb{E}[f^*(\theta_t) - \theta_t^T v_t^* | \theta_t] \\
&= \sum_{t=1}^{T} f^*(\theta_t) - \theta_t^T \bar{V} \\
&\geq \sum_{t=1}^{T} f(\bar{V}) \\
&= T \cdot OPT_f
\end{aligned}$$

Here the last step follows from the fact the $f^*(\theta) - \theta^T x$ is tangent to $f(x)$ and hence above it. Finally we can take expectation w.r.t θ_t and we get the statement of the lemma. □

Now our regret guarantee is immediate.

Theorem 6. *Algorithm 1 has expected regret atmost $O(DG\sqrt{T})$.*

Proof. Combining the previous lemma and equation 3, we get:

$$T \cdot OPT_f \leq \mathbb{E}[\sum_{t=1}^{T} f^*(\theta_t) - \theta_t^T v_t] \tag{7}$$

$$\leq \mathbb{E}[\sum_{t=1}^{T} f^*(\bar{\theta}) - \bar{\theta}^T v_t] + O(DG\sqrt{T}) \tag{8}$$

$$= T\mathbb{E} f\left(\frac{\sum_{t=1}^{T} v_t}{T}\right) + O(DG\sqrt{T}) \tag{9}$$

□

4 Algorithm for Bandit Setting

In this section we use $v_t, V_t, \hat{v}_t$ as a short hand of the reward, mean reward and empirical mean reward of the played arm I_t. We now modify our algorithm for the bandit setting. The essential idea is simple: We maintain estimates $\tilde{v}_{i,t}$ for each arm with the property:

$$f^*(\theta_t) - \theta_t^T \tilde{v}_{i,t} \geq f^*(\theta_t) - \theta_t^T V_i \text{ w.h.p.} \tag{10}$$

This is done by setting:

$$\tilde{v}_{i,t} = \hat{v}_{i,t} - sign(\theta_t) \cdot \sqrt{\frac{\ln T}{n_{i,t}}} \tag{11}$$

Here $n_{i,t}$ is the number of pulls of arm i, and $\hat{v}_{i,t}$ is the empirical mean of the arm, $\cdot$ denotes element-wise multiplication. The fact that this scheme ensures equation 10 holds follows from standard UCB arguments. The final algorithm is shown below. The regret-analysis follows from the lemmas given below.

for $t = 1 : T$ **do**
  $I_t = \arg\max_{i \in 1 \ldots N} f^*(\theta_t) - \theta_t^T \tilde{v}_{i,t}$;
  Play arm I_t, Observe $v_{I_t,t}$;
  Update $\tilde{v}_{i,t} = \hat{v}_{i,t} - sign(\theta_t) \cdot \sqrt{\frac{\ln T}{n_{i,t}}}$;
  Observe $h_t(x) = f^*(x) - x^T v_t$ and Predict θ_{t+1} using OGD.
end

Algorithm 2: Online Stochastic Convex Optimization

Lemma 7. $\sum_{t=1}^{T} f^*(\theta_t) - \theta_t^T v_t \leq \sum_{t=1}^{T} f^*(\bar{\theta}) - \bar{\theta}^T v_t + O(DG\sqrt{T}) = T f\left(\frac{\sum_{t=1}^{T} v_t}{T}\right) + O(DG\sqrt{T})$

Proof. Since the true value of v_t is available for the prediction step at time $t+1$ this bound is exactly the same as in the case of full information setting. □

Lemma 8. $T \cdot OPT_f \leq \mathbb{E}\left[\sum_{t=1}^{T} f^*(\theta_t) - \theta_t^T \tilde{v}_t\right]$

Proof. The proof is exactly the same as lemma 5 except that we begin with equation 10. □

Lemma 9. $\mathbb{E}\left[\sum_{t=1}^{T}\theta_t^T(v_t-\tilde{v}_t)\right] \leq \tilde{O}(\sqrt{NTd})$

Proof. Let $\mathcal{F}_t$ denote all the information available before selecting the arm at time t.

$$\mathbb{E}\left[\sum_{t=1}^{T}\theta_t^T(v_t-\tilde{v}_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\theta_t^T\mathbb{E}\left[v_t-\tilde{v}_t|\mathcal{F}_t\right]\right] \tag{12}$$

$$= \sqrt{\ln T}\mathbb{E}\left[\sum_{t=1}^{T}|\theta_t|^T\mathbf{1_d}\frac{1}{\sqrt{n_{I_t,t}}}\right] \qquad \text{[Plugin the value of } \tilde{v}_t\text{]} \tag{13}$$

$$\leq L\sqrt{d\ln T}\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\sqrt{n_{I_t,t}}}\right] \qquad \text{[Cauchy Schwarz]} \tag{14}$$

$$\leq L\sqrt{dNT\ln T} \tag{15}$$

□

The final regret bound is given in the following theorem:

Theorem 10. *Algorithm 2 has a regret bound of atmost* $O(DG\sqrt{T}) + \tilde{O}(\sqrt{NTd})$

Proof.

$$T\cdot OPT_f \leq \sum_{t=1}^{T}\mathbb{E}\left[f^*(\theta_t)-\theta_t^T\tilde{v}_t\right] \qquad \text{[Lemma 8]} \tag{16}$$

$$= \sum_{t=1}^{T}\mathbb{E}\left[f^*(\theta_t)-\theta_t^T v_t\right] + \tilde{O}(\sqrt{NTd}) \qquad \text{[Lemma 9]} \tag{17}$$

$$\leq Tf\left(\frac{\sum_{t=1}^{T}v_t}{T}\right) + O(DG\sqrt{T}) + \tilde{O}(\sqrt{NTd}) \qquad \text{[Lemma 7]} \tag{18}$$

□

The key references are [1],[2]

References

[1] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006. ACM, 2014.

[2] Shipra Agrawal and Nikhil R Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1405–1424. SIAM, 2015.