COMS 4995-3: Advanced Algorithms — April 17, 2017

Lecture 23 – Multiplicative Weights Update

Instructor: *Alex Andoni* — Scribes: *Alexander Yu, John Heintschel*

1 Introduction

Today's lecture continues the previous discussion on the interior point method (IPM) which aims to take a constrained optimization problem of the form:

$$\min_{x \in \mathbb{R}^n} c^T \cdot x$$
$$\text{such that } Ax \leq b$$

(let K be the feasible region) and turns it into an unconstrained optimization problem of the form:

$$\min_{x \in \mathbb{R}^n} \{\eta c^T x + F(x)\} \tag{1}$$

for $\eta > 0$ where the barrier function $F(x)$ is defined as:

$$F(x) = -\sum_{i=1}^{n} \log(b_i - A_i x)$$

2 Interior Point Method, Continued

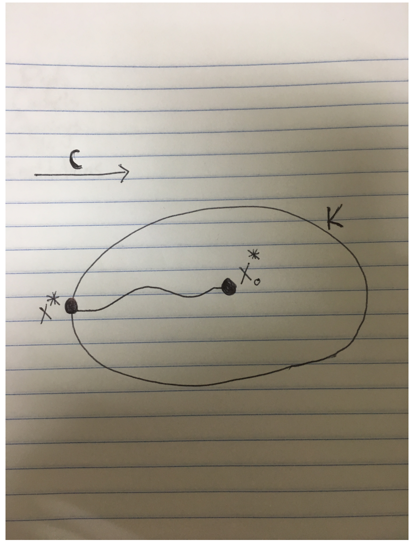
First, we observe that

$$\lim_{x \to \partial K} F(x) = \infty,$$

or in other words, $F(x)$ grows unbounded as x approaches the boundary of K (i.e. the set of $x \in K$ such that $Ax = b$), which captures the constraint that x must be in the feasible region in the original problem, i.e. that $x \in K$. Now let the optimal solution to (1) be x^*, and let

$$x^*_\eta = \arg\min\{\eta c^\top x + F(x)\}.$$

x^*_η is a continuous function which defines the central path from the analytic center of K, x^*_0, to the optimal point x^*:

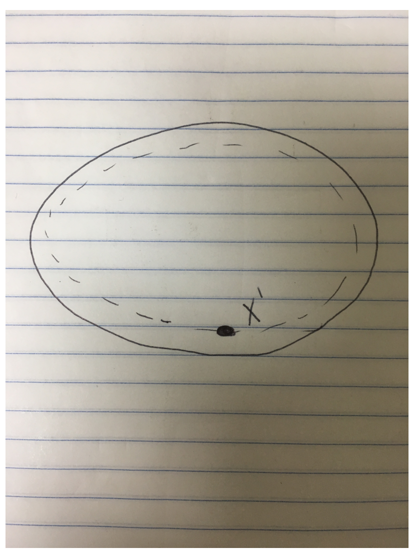


It was shown last class that $x^*_\infty = x^*$, and thus if we make η sufficiently large then we will have $x^*_\eta \to x^*$. The steps of the interior point method are as follows:

1. **Initialization:** first assume that $\mathrm{Vol}(K) > 0$. We must find some x' in the interior of K: $x' \in K \setminus \partial K$. We can find x' by solving the reduced linear program:

$$\begin{aligned} &\min t \\ \text{s.t. } & Ax \leq b + t \end{aligned}$$

   This will amount to finding a point on the dotted line, which is inside of K:



2. We now have a point on the interior of K: x'. Now we show the following claim:

   Claim 1. *$\exists \bar{c}$ such that $\exists \eta$ such that x' is the optimal value for the unconstrained optimization problem:*

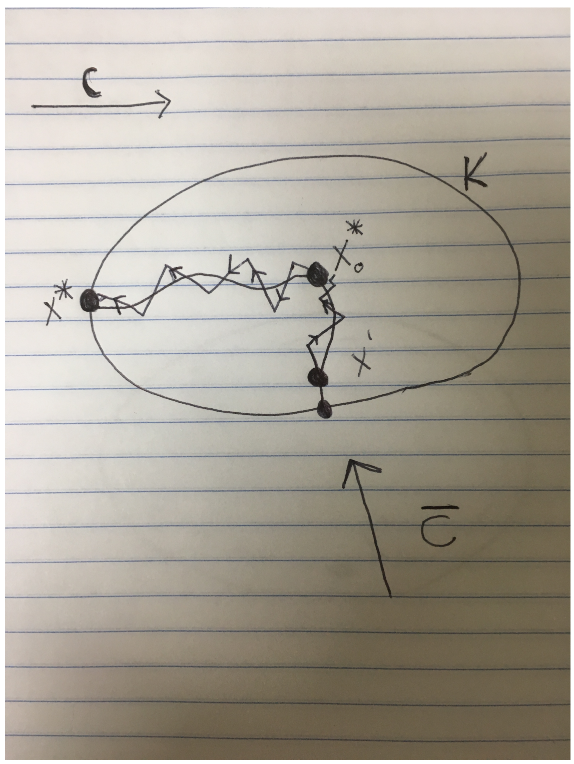
$$\min \bar{c}^\top x + F(x)$$

Proof.

- Take
$$\nabla \Big(\eta \bar{c}^T x' + F(x') \Big) = 0$$
- Then
$$\eta \bar{c}^T = -\nabla F(x')$$
and we have that $x' = x_\eta^*$ for a different objective function.

$\square$

3. Now we know that after "rotating" our feasible polytope we can say that x' lies on the central path between the analytic center and the optimum for the different objective function. Our strategy will first be to follow along the path using Newton steps back to the analytic center, using smaller and smaller η, and then travel from the analytic center to the original x^*. Graphically, this looks like:



We start at x', follow along the jagged path to the analytic center, and then travel to x^*. This will converge to the solution of our original L.P. problem.

Remark. As motivation for why IPM might be used as a method, note that in general, we wish to optimize some convex function $f(x)$. We have two choices:

1. Gradient Descent. This has first-order convergence on the order of $\frac{1}{\epsilon}$ (too slow).
2. Newton's Method. This converges in time dependent on $\log\log\frac{1}{\epsilon}$, which is much faster, but requires that the starting point is "close" to x^*. Thus we use IPM to get a close starting point.

3 Multiplicative Weights

Consider the problem of deciding whether a given stock will rise or fall in price over the course of a day. We have access to n experts who tell us over the course of T days (on every day) whether they believe the price will rise or fall. Then for $i = 1, 2, \ldots, n$, let:

$$f_i^t = \begin{cases} 1 & \text{if expert } i \text{ at day } t \text{ is wrong,} \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, define

$$\begin{aligned} m_i^t &= \text{the number of errors expert } i \text{ did in the first } t \text{ days,} \\ M^t &= \text{number of errors made by us in the first } t \text{ days.} \end{aligned}$$

Goal: make M^t as close as possible to the number of errors of the best expert; in other words, make M^t as close as possible to $\min_i m_i^t$.

The complication is that we cannot make any assumptions about the behavior of the experts, such as assumptions about their accuracy rate, correlation, etc., and may even be adversarial in nature. This thus invalidates some simple procedures that may come to mind, such as the following non-solutions:

- Taking the majority vote of the experts
  - **Counterexample:** Take $n = 3$, and suppose that for each day, the correct action is to sell. However, experts 1 and 2 always recommend to buy. Then for each day, we thus choose to buy and make an error for each day. As a result, for each t we get that $M^t = t$ (i.e. the worst possible value).
- Following the expert that was right the previous day
  - **Counterexample:** Take $n = 2$, and suppose that for each day, the correct action is to sell. However, expert 1 recommends to buy on even days and sell on odd days, and expert 2 recommends to sell on even days and buy on odd days. Then for each $t \geq 1$, we always will choose to buy, and thus make an error. As a result, for each $t \geq 1$ we get that $M^t \geq t - 1$ (depending on whether or not we choose the correct action on $t = 0$).

3.1 Weighted Majority Algorithm

Idea: use all past history on experts to determine our decision. In particular, penalize experts that are tend to be wrong, and try to listen more to experts that tend to be right.

The above idea then leads to the following algorithm, which we call the **Weighted Majority Algorithm**:

1. For $i = 1, 2, \ldots, n$, for day $t = 0$ assign expert i a weight of $w_i^0 = 1$.
2. For each $t \geq 0$, set $w_i^{t+1} = w_i^t(1 - \varepsilon f_i^t)$, where ε is a parameter to be determined. In other words, decrease the weight of each expert i by a factor of $1 - \varepsilon$ if it was incorrect on day t, and otherwise do not change the weight.
3. The decision for each day t is a weighted majority: take the sum of the weights of the experts recommending to sell, and the sum of the weights of the experts recommending to buy, and choose the option with the higher weighted sum.

3.2 Analysis of the Weighted Majority Algorithm

It turns out that the Weighted Majority Algorithm, for appropriate choices of ε, yields a decent value of M^T relative to the best expert. In particular, we have the following theorem:

Theorem 2. *$\forall i = 1, 2, \ldots, n$, $\forall \varepsilon \in (0, 1/2)$, we have that for any T, the Weighted Majority Algorithm yields that*

$$M^T \leq 2(1+\varepsilon)m_i^T + \frac{2\ln n}{\varepsilon},$$

or equivalently, that

$$M^T \leq 2(1+\varepsilon)\min_i m_i^T + \frac{2\ln n}{\varepsilon}.$$

Proof. For $t = 0, \ldots, T$, define

$$\Phi_t = \sum_{i=1}^{n} w_i^t.$$

Note that $\Phi_0 = n$. Next, we prove the following lemmas:

Lemma. *For $t \geq 0$, we have that*

$$\Phi_{t+1} \geq w_i^{t+1} = (1-\varepsilon)^{m_i^t}. \tag{2}$$

Proof. The inequality portion of (2) follows immediately from our definition of Φ_t, since we have that $w_i^t \geq 0$ for all i. Moreover, note that

$$w_i^{t+1} = (1-\varepsilon)^{m_i^t},$$

since the weight of expert i at the beginning of day $t+1$ will simply be $w_i^0 = 1$ multiplied by $1 - \varepsilon$ precisely m_i^t times, since the weight of expert i is only decreased whenever expert i makes an error, which is precisely what m_i^t counts up until day t. Thus, (2) holds, as desired. □

Lemma. *For $t \geq 0$, we have that*

$$\Phi_{t+1} \leq n\left(1 - \frac{\varepsilon}{2}\right)^{M^t}.$$

Proof. For each $t \geq 0$, note that if we choose the correct action on day t, then clearly $\Phi_{t+1} \leq \Phi_t$, since by definition we have that $w_i^{t+1} \leq w_i^t$, with equality iff expert i was correct. This is tight, since it may be possible that all experts choose correctly on day t. Alternatively, suppose we chose wrong. Then by our algorithm, this means that the weighted majority chose wrong. Suppose the set of experts that chose wrong was $A \subset \{1, 2, \ldots, n\}$. Then since it was the weighted majority, we have that

$$\sum_{i \in A} w_i^t \geq \frac{1}{2} \Phi_t. \tag{3}$$

Moreover, note that for each expert $i \in A$, since it chose wrong, we decrease its weight, and for each expert $i \notin A$, we keep its weight the same. Then it follows that

$$\begin{aligned}
\Phi_{t+1} &= \sum_{i=1}^{n} w_i^{t+1} \\
&= \sum_{i \in A} w_i^{t+1} + \sum_{i \notin A} w_i^{t+1} \\
&= \sum_{i \in A} w_i^t (1 - \varepsilon) + \sum_{i \notin A} w_i^t \\
&= \sum_{i=1}^{n} w_i^t - \varepsilon \sum_{i \in A} w_i^t \\
&= \Phi_t - \varepsilon \sum_{i \in A} w_i^t \\
&\overset{(3)}{\leq} \Phi_t - \frac{\varepsilon}{2} \Phi_t \\
&= \Phi_t \left(1 - \frac{\varepsilon}{2}\right).
\end{aligned}$$

In summary, since the above inequality holds for each day where we choose incorrectly, M^t counts precisely all such incorrect choices up until day t, and $\Phi_{t+1} \leq \Phi_t$ even when we choose correctly, it easily follows by induction that

$$\Phi_{t+1} \leq \Phi_0 \left(1 - \frac{\varepsilon}{2}\right)^{M^t} = n \left(1 - \frac{\varepsilon}{2}\right)^{M^t},$$

as desired. □

Combining the two lemmas above yields that for each $t \geq 0$, we have that

$$(1 - \varepsilon)^{m_i^t} \leq \Phi_{t+1} \leq n \left(1 - \frac{\varepsilon}{2}\right)^{M^t} \leq n e^{-\varepsilon/2 \cdot M^t},$$

where the last inequality comes from the fact that $1 - x \leq e^{-x}$ for all x. Now noting that

$$\ln(1 - \varepsilon) \geq -\varepsilon - \varepsilon^2, \quad \varepsilon \in (0, 1/2), \tag{4}$$

we get that

$$\begin{aligned}
(1 - \varepsilon)^{m_i^t} \leq n e^{-\varepsilon/2 \cdot M^t} &\implies m_i^t \ln(1 - \varepsilon) \leq \ln n - \frac{\varepsilon}{2} M^t \\
&\overset{(4)}{\implies} -m_i^t \varepsilon (1 + \varepsilon) \leq \ln n - \frac{\varepsilon}{2} M^t
\end{aligned}$$

$$\implies M_t \leq 2(1+\varepsilon)m_i^t + \frac{2\ln n}{\varepsilon}.$$

Since this holds for any $t \geq 0$, simply setting $t = T$ yields the desired conclusion. □

3.3 Multiplicative Weights Update

While the above analysis of the Weighted Majority Algorithm yields a nice guarantee that M_t is within a constant multiple of m_i^t (ignoring the $\frac{2\ln n}{\varepsilon}$ term, which is a fixed constant, given that we fix the number of experts n and the penalization parameter ε beforehand), this constant is still about $2(1+\varepsilon) \approx 2$.

Now consider the more general situation where we can actually now assume that for each day t and expert i, that f_i^t is no longer simply an indicator variable, but can take a range of values in $[-1,1]$. Another way to view it is generalizing the loss function of expert i on day t from simply being a 0-1 loss function to any loss function with a range in $[-1,1]$; f_i^t now represents a measure of how much was lost on day t by expert i.

Then to reduce the factor of 2 introduced by the Weighted Majority Algorithm, we can improve upon the algorithm by adding a randomization aspect, which leads to an algorithm which we call **Multiplicative Weights Update**. In this algorithm, we perform the same initialization and update of weights, but instead of basing our decision off of the weighted sum, we perform the following process for each day t:

1. For each expert i, compute
$$p_i^t = \frac{w_i^t}{\sum_{j=1}^n w_j^t}.$$

2. Choose an expert i, such that each expert i is chosen with probability p_i^t.

3. The decision for day t is precisely the recommendation of expert i.

3.4 Analysis of Multiplicative Weights Update

Within the context of this algorithm, due to the randomization aspect of choosing an arbitrary expert, we thus redefine our notion of the number of our mistakes M^T. More precisely for $t = 1, \ldots, T$, define

$$p^t = (p_1^t, \ldots, p_n^t),$$
$$f^t = (f_1^t, \ldots, f_n^t).$$

Then we define M^T as
$$M^T = \sum_{t=1}^T \langle p^t, f^t \rangle = \sum_{t=1}^T \sum_{i=1}^n p_i^t f_i^t.$$

Previously, M^T was simply defined as the number of mistakes that we had made up until day T. This can be viewed as the result of a loss function, where we applied a 0-1 loss function to our algorithm for each day. Here, however, one perspective is that we are redefining our loss to be the expected number of errors that our algorithm makes until day T.

Alternatively, another perspective on our algorithm is that we actually apportion our money among all the experts proportionally according to the weights (or equivalently, the probabilities), and attribute each expert to handle that portion of the money based on their own recommendations. In this case, M^T

is actually the total loss of money among all of the experts over all the days and is deterministic, rather than simply being an expectation of total loss over randomly chosen experts for each day.

Moreover, since f_i^t is no longer an indicator variable, we must also redefine our notion of loss for each expert i. Previously, m_i^T referred to the total number of mistakes that expert i made until day T, which similarly to our old definition of M^T could be viewed as the sum of 0-1 loss functions. Now, we extend the same notion algebraically by simply using our new loss functions, and defining

$$m_i^T = \sum_{t=1}^{T} f_i^t.$$

With these notions of loss for both us and the experts redefined, it turns out that the Multiplicative Weights Update algorithm yields the following improvement:

Theorem 3. *$\forall i = 1, 2, \ldots, n$, $\forall \varepsilon \in (0, 1/2)$, we have that for any T, the Multiplicative Weights Update algorithm yields that*

$$M_T \leq (1 + \varepsilon) m_i^T + \frac{\ln n}{\varepsilon},$$

or equivalently, that

$$M_T \leq (1 + \varepsilon) \min_i m_i^T + \frac{\ln n}{\varepsilon}.$$