CSE599s, Spring 2012, Online Learning Lecture 6 - 04/12/2012

The Online Gradient Descent with adaptive learning rate

Lecturer: Brendan McMahan *Scribe: Yanping Huang*

1 The Online Gradient Descent Algorithm

In the previous lecture, Zinkevich's online gradient descent [1] algorithm was presented:

```
ONLINE GRADIENT DESCENT (OGD).
Inputs: convex feasible set $\mathcal{W} \in \mathbb{R}^n$, non-increasing step sizes $\eta_1, \eta_2, \ldots \geq 0$, initial $\mathbf{w}_0 \in \mathcal{W}$
for $t = 1, 2, \ldots,$ do
  $\mathbf{w}_t = \Pi_{\mathcal{W}}(\mathbf{w}_{t-1} - \eta_t \mathbf{g}_t)$, where $\mathbf{g}_t \in \partial f_t(\mathbf{w}_t)$
end for
Here $\Pi_{\mathcal{W}}$ denotes the projection onto nearest point in $\mathcal{W}$, $\Pi_{\mathcal{W}}(\mathbf{w}) = \arg\min_{\mathbf{w}' \in \mathcal{W}} \|\mathbf{w} - \mathbf{w}'\|$.
```

If we use a feasible set where $\|\mathbf{w}\| \leq R$, we showed a general bound for this algorithm of

$$\text{Regret} \leq \frac{2R^2}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta g_t^2. \tag{1}$$

In the previous lecture, assuming $f_1, f_2, \ldots, f_T$ are G-Lipschitz, we let $\eta_t = \frac{R\sqrt{2}}{G\sqrt{t}}$, and showed the above bound reduces to

$$\text{Regret} \leq 2\sqrt{2}RG\sqrt{T}. \tag{2}$$

Note that this regret bound is the bound for infinite horizon problems, *i.e.*, the algorithm needs not know the total number of iterations T in advance. The bound holds on the regret up through round T for all $T \geq 1$. The regret bound for the corresponding finite horizon problems (which holds only for a fixed T, which we need to know in advance) can be shown to be $2RG\sqrt{T}$.

Some comments on OGD algorithm:

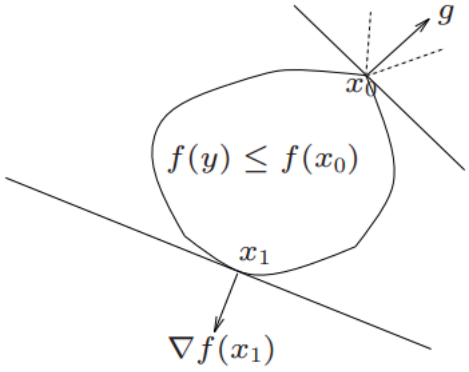
- Gradient descent problems with smaller feasible sets $\mathcal{W}$ are easier.
  This can be easily shown by the regret bound in Eq. (2).

- The potential function may not be monotonically increasing or decreasing, depending on the choice of $\mathbf{w}^\star$. Recall $\Phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}^\star\|^2$ and let the projection operator $\Pi_{\mathcal{W}}$ be the identity operator, that is $\mathcal{W} = \mathbb{R}^n$. We have the below recursive relationship:

$$\begin{aligned}\Phi(\mathbf{w}_{t+1}) &= \frac{1}{2}\|\mathbf{w}_t - \mathbf{w}^\star\|^2 - \eta_t \mathbf{g}_t^T(\mathbf{w}_t - \mathbf{w}^\star) + \frac{1}{2}\eta_t^2\|g_t\|^2 \\ &= \Phi(\mathbf{w}_t) + \frac{1}{2}\eta_t^2\|g_t\|^2 - \eta_t \mathbf{g}_t^T(\mathbf{w}_t - \mathbf{w}^\star).\end{aligned}$$

If $f_t(\mathbf{w}^\star) \leq f_t(\mathbf{w}_t)$, we have $\mathbf{g}_t^T(\mathbf{w}^\star - \mathbf{w}_t) \leq 0$ from Lemma 1 (below), then $\Phi(\mathbf{w}_{t+1})$ will be smaller than $\Phi(\mathbf{w}_t)$ if we choose a small enough learning rate. On the other hand, if $f_t(\mathbf{w}^\star)$ is not a desired point, *i.e.*, $f_t(\mathbf{w}^\star) > f_t(\mathbf{w}_t)$, we may have $\mathbf{g}_t^T(\mathbf{w}^\star - \mathbf{w}_t) > 0$. In this case $\Phi(\mathbf{w}_{t+1}) > \Phi(\mathbf{w}_t)$.

Lemma 1. *Let* $\mathbf{g}$ *be a sub-gradient of* f *at* $\mathbf{x}$. *If* $f(\mathbf{y}) \leq f(\mathbf{x})$ *then* $\mathbf{g}^T(\mathbf{y} - \mathbf{x}) \leq 0$ *(immediate from* $f(\mathbf{y}) \geq f(\mathbf{x}) + g_t^T(\mathbf{y} - \mathbf{x})$*). Thus, nonzero subgradients of* f *at* x *define supporting hyperplanes to sub-level set* $\{\mathbf{y} | f(\mathbf{y}) \leq f(\mathbf{x})\}$ *(see picture).*



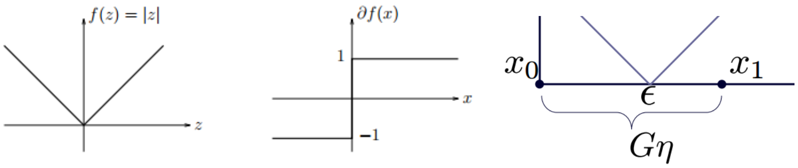
- In a later lecture, we will analyze the Follow-the-Regularized-Leader (FTRL) algorithm with adaptive regularization, where on each round we use a total amount of regularization like $R(\mathbf{w}) \approx \sqrt{t}\|\mathbf{w}\|^2$.

2 Lower Bounds for OGD

In this section, we would like to show that the regret bound for OGD is tight by constructing problems whose regret is (up to constant factors) as large as the regret bound for OGD. Note: In this section, we use x's instead of w's.

Example 1 Large learning rates are bad. First we construct a online convex optimization problem with loss function on every round is $f_t(x) = G|x - x^\star|$, with $x \in \mathbb{R}$ and $G \in \mathbb{R}$ a constant. The corresponding subgradient can be written as $\text{sign}(x - x^\star)$, where the sign function $\text{sign}(x) = 1$ for $x > 0$ and $\text{sign}(x) = -1$ for $x < 0$. The OGD update rule $x_{t+1} = x_t - \text{sign}(x_t - x^\star)\eta G$ will then make the sequence $\{x_t\}$ oscillate around the optimal point $x^\star$. Suppose at time t, $x^\star - x_t = \epsilon$ with $0 < \epsilon < G\eta$. Then, we will have $x_{t+1} = x_0 + G\eta$, so x_t and x_{t+1} are in the opposite sides of $x^\star$. After the next update, we will have $x_{t+2} = x_{t+1} - G\eta = x_t$, and so the oscillation continues. The resulting regret will be

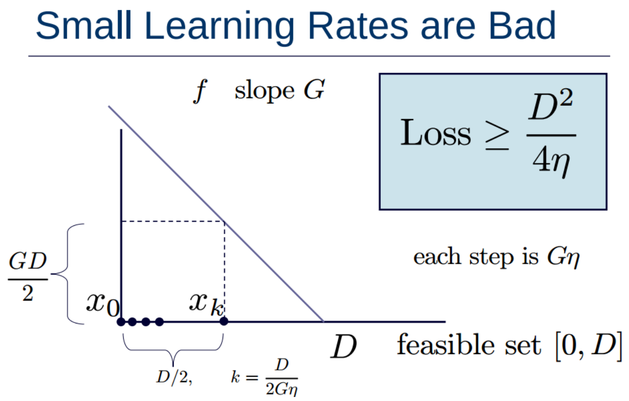
$$\text{Regret} = \frac{T}{2}G\epsilon + \frac{T}{2}G(G\eta - \epsilon) = \frac{T}{2}G^2\eta.$$



The $\frac{T}{2}G\epsilon$ term counts the $T/2$ rounds where we are at the first point (without loss of generality, $x_0, x_2, \ldots, x_t, x_{t+2}, \ldots$), when we are a distance ϵ from $x^\star$. The term $\frac{T}{2}G(G\eta - \epsilon)$ counts the regret on the remaining $T/2$ rounds when we are a distance $G\eta - \epsilon$ from $x^\star$.

Example 2 Small learning rates are bad, too. Let $x \in [0, D]$, and $f_t(x) = G(D - x)$. The OGD update rule $x_{t+1} = x_t + G\eta$ will generate a sequence of $\{x_t\} = \{x_0, x_0 + G\eta, x_0 + 2G\eta, \ldots,\}$. Let $x_0 = 0$, after $K = \frac{D}{2G\eta}$ steps, the total regret will be

$$\text{Regret} = \sum_{t=0}^{k} G(D - tG\eta) \geq \frac{GD}{2}\frac{D}{2G\eta} = \frac{D^2}{4\eta}.$$



Small Learning Rates are Bad

$$\text{Loss} \geq \frac{D^2}{4\eta}$$

Using previous two constructions, for any learning rate, the adversary can choose a one dimensional problem where regret is at least $\max\{\frac{D^2}{4\eta}, G^2\eta\frac{T}{2}\}$. In comparison, the regret bound for OGD with feasible region $[0, D]$ and maximum gradient G has the form of $\frac{D^2}{2\eta} + G^2\eta\frac{T}{2}$. This shows the regret bound for OGD is tight.

3 OGD with adaptive learning rate.

Our goal now will be to study Eq. (1), and to derive better learning rate schedules which produce lower regret bounds. We begin with a motivating example that shows when this may be possible.

Suppose in the online optimization game (in one dimension), the adversary plays a sequence like

$$g_t = \{0, 0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, 1, 0, \ldots\},$$

$t = 1, \ldots, T$. For example, let $T = 10^{10}$ but only 10^4 of g_t are non-zero. The OGD with $\eta_t \approx 1/\sqrt{t}$ will have a step size decreasing on each round. This choice of step size will have a regret bound of order $\sqrt{T} = 10^5$. Alternatively, if we update the step size only when $g_t \neq 0$, we will have a much lower regret bound of order $\sqrt{10^4} = 10^2$. The key is that we can safely ignore rounds when $g_t = 0$, because whatever w_t we choose will incur the same loss as any $w^\star$, so our regret will be the same as the OGD algorithm that is only updated when there are non-zero gradients.

To deal with the case where the adversary replaces 0s with some infinitesimal number $\epsilon \simeq 0$, a more general step size updating rule is needed. Suppose we know the sequence $\{g_t\}_{t=1,\ldots,T}$ in advance, the optimal fixed learning rate $\hat{\eta} = \frac{2R}{\sqrt{\sum_{t=1}^T g_t^2}}$ will have a regret at most $2R\sqrt{\sum_{t=1}^T g_t^2}$. This can be derived by taking derivatives to optimize for the best fixed rate, as we have done several times already.

But, we can do almost as well without knowing any of the g_t in advance! We use the adaptive global learning rate:

$$\eta_t = \frac{R\sqrt{2}}{\sqrt{\sum_{s=1}^t g_s^2}} \tag{3}$$

The corresponding regret is

$$\text{Regret} \leq 2\sqrt{2}R\sqrt{\sum_{t=1}^T g_t^2} \tag{4}$$

It is not obvious that plugging the learning rate from Eq. (3) into the bound of Eq. (1) gives a bound like Eq. (4). Proving this requires a technical lemma, see for example [2].

We apply the above algorithm to the following online prediction problem:

Example 3 Online Prediction Problem. Let $\{\mathbf{x}_t, y_t\}$ be a sequence of learning examples where $\mathbf{x}_t$ represents the feature vector and y_t represents the label. $\mathbf{x}_t \in \mathbb{R}^n$, $y_t \in \{0, 1\}$. We will make predictions using a **generalized linear model**, where $\hat{y}_t = \sigma(\mathbf{w}_t, \mathbf{x}_t)$ based on $\mathbf{x}_t$ and parameters $\mathbf{w}_t$. The nonlinear function $\sigma(\cdot)$ maps from $\mathbb{R}$ to $[0, 1]$. For example $\sigma(\cdot)$ can be a sigmoid function $\sigma(x) = \frac{e^x}{1+e^x}$. Then this online prediction problem can be viewed as logistic regression problem. At each round t, the player receives $\mathbf{x}_t$, makes a prediction $\hat{y}_t$ by choosing a $\mathbf{w}_t$, and suffers a loss $f_t(\mathbf{w}_t) = \ell(\mathbf{w}_t \cdot \mathbf{x}_t, y_t)$. For generalized linear models, ℓ is usually chosen so that $g_t = \nabla f_t(w_t) = (\sigma(\mathbf{w}_t \cdot \mathbf{x}_t) - y_t)\mathbf{x}_t = (\hat{y}_t - y_t)\mathbf{x}_t$.

In problems like document classification, the feature vector $\mathbf{x}_t$ is usually very sparse. For example, $x_{t,i}$ may represent whether a word i appears in a document t or not. While x_t is in a very high n-dimensional space (n is the number of words in a dictionary), typically most $x_{t,i}$ are zero (since most documents have a relatively small number of distinct words).

If we choose a feasible set $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^n \mid w_i \in [-B_i, B_i]\}$ for constants $B_i \in \mathbb{R}$, we can apply OGD with the learning rate from Eq. (3) on a per-coordinate basis:

$$w_{t+1,i} = \Pi_{[-B_i,B_i]}(w_{t,i} - \eta_{t,i} g_{t,i})$$
$$\text{where} \quad \eta_{t,i} = \frac{\sqrt{2}B_i}{\sqrt{\sum_{s=1}^{t} g_{s,i}^2}}.$$

The regret can be shown to be no more than $\sum_{i=1}^{n} 2B_i \sqrt{2\sum_{t=1}^{T} g_{t,i}^2}$. Writing $\mathbf{B} = (B_1, B_2, \ldots, B_n)$ and $\mathbf{g}_t = (\ldots, \sqrt{\sum_{t=1}^{T} g_{t,i}^2}, \ldots)$, We have

$$\text{Regret} \le 2\sqrt{2}\mathbf{B} \cdot \mathbf{g}_t \le 2\sqrt{2}\|\mathbf{B}\|\|\mathbf{g}_t\| = 2\sqrt{2}R\sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|^2},$$

and so the adaptive per-coordinate learning rate gives a bound as least as good as the adaptive global rate, Eq. (3).

References

[1] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent", in *International Conference on Machine Learning*, 2003.

[2] M. Streeter and H.B. McMahan, "Less Regret via Online Conditioning", `http://arxiv.org/abs/1002.4862`, 2010.