Less Regret via Online Conditioning

Sampad Kumar Kar

Chennai Mathematical Institute

December 20, 2023

Motivation

- *Less Regret via Online Conditioning* by Matthew Streeter and H. Brendan McMahan (2010).
- We analyze and evaluate an **Online Gradient Descent** algorithm with adaptive per coordinate adjustment of learning rates.
- This leads to regret bounds that are stronger than those of standard online gradient descent for general online convex optimization problems. This is also evident empirically.

Formulation

- Recall definition of regret in OCO setting:

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x)$$

- Simplest algorithm that applies to the most general setting of OCO is **Online Gradient Descent** (OGD) algorithm by Zinkevich (2003).

Online Gradient Descent

Online Gradient Descent Algorithm

- Input: Convex Set $\mathcal{K}$, T, $x_1 \in \mathcal{K}$, stepsizes $\{\eta_t\}$
- for $t = 1$ to T do:
    - Play x_t and observe the cost $f_t(x_t)$
    - Update and Project:

$$\begin{aligned} y_{t+1} &= x_t - \eta_t g_t \\ x_{t+1} &= \Pi_{\mathcal{K}}(y_{t+1}) \end{aligned}$$

    - Here g_t is a subgradient of f_t at x_t.

OGD Regret Bounds

Theorem 1

OGD with stepsize $\{\eta_t = \frac{D}{G\sqrt{t}}, t \in [T]\}$ guarantees the following for all $T \geq 1$:

$$\mathrm{Regret}_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x) \leq \frac{3}{2} GD\sqrt{T}$$

OGD Regret Bounds

Theorem 1

OGD with stepsize $\{\eta_t = \frac{D}{G\sqrt{t}}, t \in [T]\}$ guarantees the following for all $T \geq 1$:

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x) \leq \frac{3}{2} GD\sqrt{T}$$

Proof:

$f_t(x_t) - f_t(x^*) \leq \nabla^T f_t(x_t)(x_t - x^*)$; By Convexity (*)

$\|x_{t+1} - x^*\|^2 \leq \|y_{t+1} - x^*\|^2$; By Pythagoras Theorem

$$\nabla^T f(x_t)(x_t - x^*) \leq \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta_t} + \frac{\eta_t G^2}{2};$$

By substituting y_{t+1} from OGD (**)

Proof

$$f_t(x_t) - f_t(x^*) \leq \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta_t} + \frac{\eta_t G^2}{2};$$

using (*) and (**)

$$\text{Regret}_T \leq \sum_{t=1}^{T} \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta_t} + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t;$$

Summing above from $t = 1$ to T

$$\text{Regret}_T \leq \frac{D^2}{2\eta_T} + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t; (\dagger)$$

$$\text{Regret}_T \leq \frac{3}{2}GD\sqrt{T};$$

Assuming G- Lipchitz, D- Diameter and $\frac{1}{\eta_0} = 0$ and $\eta_t = \frac{D}{G\sqrt{t}}$

A Motivating Application

- **Problem:** Predicting the probability that a user will click on an ad when it is shown alongside search results for a particular query, using a **Generalized Linear Model** (GLM):
  - On round t, Algorithm predicts $p_t(x_t) = l(x_t.\theta_t)$
  - $x_t, \theta_t \in \mathbb{R}^n$: vector of weights, features, l: link function
  - Ex: $l(\alpha) = \frac{1}{1+\exp -\alpha}$, $l(\alpha) = \alpha$
  - Algorithm incurs loss, which is some convex function of p_t; Ex: cross entropy loss, sum square loss
  - Played in OGD setting: $x_{t+1} = \Pi_{\mathcal{K}}(x_t - \eta_t g_t)$; g_t being a subgradient of f_t at x_t
- Single Ad System, we wish to predict it's click-through rate on different queries.
- On a large search system, a popular query will occur orders of magnitude more often than a rare query:
  - rare queries: need larger learning rates
  - popular queries: smaller learning rates

Tradeoffs in 1-Dimension: Global Learning Rate

- Consider $\mathcal{F} = [0, D]$.
- When η is too **large**:
  - Let $f_t(x) = G|x - \epsilon|$. Then:

$$\nabla f_t(x) = \begin{cases} -G & x \in [0, \epsilon] \\ G & x \in (\epsilon, D] \end{cases}$$

  - If $x_1 = 0$, then OGD plays[1] $x_t = 0$ on odd rounds and $x_t = G\eta$ on even rounds. Here, $x^* = \epsilon$.
  - This incurs: $\text{Regret}_T = \frac{T}{2} G\epsilon + \frac{T}{2} G(G\eta - \epsilon) = \frac{T}{2} G^2 \eta$.
  - Note: The regret is not sublinear.

[1] Assuming $\epsilon < G\eta < D$

Tradeoffs in 1-Dimension: Global Learning Rate

- Consider $\mathcal{F} = [0, D]$.
- When η is too **small**:
  - Let $f_t(x) = -Gx$. Then $\nabla f_t(x) = -G$
  - If $x_1 = 0$, then OGD plays $x_t = \min\{D, (t-1)G\eta\}$. Here, $x^* = D$.
  - Since we are assuming η to be small[2]:
    - For first $\frac{D}{2G\eta}$ rounds, per round regret is atleast $\frac{GD}{2}$.
    - Also[3], $\text{Regret}_T \geq \frac{GD}{2} \times \frac{D}{2G\eta} = \frac{D^2}{4\eta}$

[2] For $t < \frac{D}{2G\eta}$, we have $x_t \leq \frac{D}{2}$

[3] Assuming $\frac{D}{2G\eta} < T$

Tradeoffs in 1-Dimension: Global Learning Rate

- Consider $\mathcal{F} = [0, D]$.
- Based on the above examples, thus for any choice of η, there exists a problem where (‡):

$$\max\{\frac{D^2}{4\eta}, \frac{T}{2}G^2\eta\} \leq \text{Regret}_T \leq \frac{D^2}{2\eta} + \frac{T}{2}G^2\eta$$

  - Upper Bound is adopted from †[4]. By setting $\eta = \frac{D}{G\sqrt{T}}$ (which minimizes the upper bound), we can minimize the worst case regret upto a constant factor[5]:
    - Optimal choice of η is proportional to $\frac{D}{G}$
    - Feasible set is large and gradients are small: large learning rates
    - Feasible set is small and gradients are large: small learning rates

[4] Substitute a global $\eta_t = \eta$

[5] This leads to a regret bound of $DG\sqrt{T}$

A Toy Example against Global Learning Rates

Theorem 2

There exists a family of online convex optimization problems, parameterized by T, where gradient descent with a *non-increasing* **global learning rate** incurs regret atleast $\Omega(T^{\frac{2}{3}})$, whereas gradient descent with an appropriate **per-coordinate learning rate** has a regret bound of $\mathcal{O}(\sqrt{T})$

A Toy Example against Global Learning Rates

Theorem 2

There exists a family of online convex optimization problems, parameterized by T, where gradient descent with a *non-increasing* **global learning rate** incurs regret atleast $\Omega(T^{\frac{2}{3}})$, whereas gradient descent with an appropriate **per-coordinate learning rate** has a regret bound of $\mathcal{O}(\sqrt{T})$. [a]

[a] This does not contradict the previously stated bound of $\mathcal{O}(GD\sqrt{T})$, as in this family of problems $D = T^{\frac{1}{6}}$ and $G = 1$.

Proof: We interleave the instances of the two classes of 1-dimensional subproblems discussed previously, by setting $G = 1$ on the feasible set $[0, 1]$. Here $\mathcal{F} = [0, 1]^n$, n is the dimension.

- We have the first subproblem of first type lasting for T_0 rounds.
- Then we have C subproblems of second type, each lasting for T_1 rounds.

Proof (Global Learning Rate)

Here is the loss function[6]:

$$f_t(x_t) = \begin{cases} |x_{t,1} - \epsilon| & t \leq T_0 \\ -x_{t,j} & t > T_0, \text{where } j = 1 + \lceil \frac{t - T_0}{T_1} \rceil \end{cases}$$

- Each round depends on exactly 1 coordinate[7].
- As observed from the 1-dimensional examples, we can easily show that $x^* = (\epsilon, 1, \ldots, 1, *, \ldots, *)$ [8]. Using this and the bounds obtained from ‡ [9]:
  - $\text{Regret}_T \geq \frac{T_0}{2}\eta + C \min\{\frac{1}{4\eta}, \frac{T_1}{2}\}$
  - Set $C = T_1 = T_0^{\frac{2}{3}}$ and assume $T_1 \leq \frac{1}{2\eta}$
  - Simple minimization over η shows that the sum is $\Omega(T_0^{\frac{2}{3}})$, which is also $\Omega(T^{\frac{2}{3}})$ as $T = T_0 + T_0^{\frac{2}{3}} \leq 2T_0$

[6] There was a slight typo here in the original paper

[7] So, exactly 1 component of gradient is non-zero every round

[8] This is ϵ followed by C 1's, with the remaining elements being irrelevant

[9] For the 1-D subproblems: $\max\{\frac{D^2}{4\eta}, \frac{T}{2}G^2\eta\} \leq \text{Regret}_T \leq \frac{D^2}{2\eta} + \frac{T}{2}G^2\eta$

Proof (Per-Coordinate Learning Rate)

So, for gradient descent with **global learning rate**, we obtain a regret lowerbound of $\Omega(T^{\frac{2}{3}})$.

- We *minimize* the regret upper bounds on a per-coordinate basis. We set the learning rates[10] in the following manner:
  - $\eta_t = \sqrt{\frac{1}{T_0}}$ for first T_0 rounds and $\eta_t = \sqrt{\frac{1}{T_1}}$ for the remaining $CT_1 + k$ rounds[11].
  - At this learning rate, we accumulate regret upper bounded by $\sqrt{T}$ for each subproblem of T rounds[12].
  - Using the $\text{Regret}_T \leq \sqrt{T_0} + C\sqrt{T_1} = 2\sqrt{T_0}$, which means $\text{Regret}_T \in \mathcal{O}(\sqrt{T_0})$ or $\mathcal{O}(\sqrt{T})$.

So, for gradient descent with **per-coordinate learning rate**, we obtain a regret upperbound of $\mathcal{O}(\sqrt{T})$.

[10] This is what makes this example a per-coordinate learning rate example, because at every iteration exactly one coordinate is meaningfully affected

[11] This is justified by the fact that we are minimizing the upper bound of $\frac{D^2}{2\eta} + \frac{T}{2}G^2\eta$ on a per coordinate basis, which gives $\eta^* = \sqrt{\frac{1}{T}}$ for T rounds

[12] Assuming $D, G = 1$ and $C = T_1 = T_0^{\frac{2}{3}}$

Improved Global Learning Rate

As shown during the proof of **Theorem 1**, by Zinkevich (2003) showing the regret upper bound of *OGD* at general OCO setting, we have the following:

$$\text{Regret}_T \leq \mathcal{B}(\eta_1, \eta_2, \ldots, \eta_T) := \frac{D^2}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2 \eta_t$$

Although the gradients $g_1, g_2 \ldots, g_T$ are not known in advance, we can come within a factor of $\sqrt{2}$ on the optimal bound, i.e. $R_{\min}$.[13]

Theorem 3

Setting $\eta_t = \frac{D}{\sqrt{2\sum_{i=1}^{t} \|g_i\|^2}}$ yields regret upper bound of $D\sqrt{2\sum_{t=1}^{T} \|g_t\|^2} = \sqrt{2}R_{\min}$.

[13] $R_{\min} := \min_{\eta_1 \leq \eta_2 \leq \cdots \leq \eta_T} \mathcal{B}(\eta_1, \eta_2, \ldots, \eta_T) = D\sqrt{\sum_{t=1}^{T} \|g_t\|^2}$

Proof

Theorem 3

Setting $\eta_t = \frac{D}{\sqrt{2\sum_{i=1}^{t}\|g_i\|^2}}$ yields regret upper bound of $D\sqrt{2\sum_{t=1}^{T}\|g_t\|^2} = \sqrt{2}R_{\min}$.

Proof: We first compute the value of $R_{\min}$. As constrained by the proof of **Theorem 1**, we only consider non-increasing sequences of $\{\eta_t\}$[14].

- This means the bound can be minimized by a constant learning rate η^*. Simple gradient minimization shows $\eta^* = \frac{D}{\sqrt{2\sum_{i=1}^{T}\|g_i\|^2}}$, which gives $R_{\min} = D\sqrt{\sum_{t=1}^{T}\|g_t\|^2}$.

Plugging this choice of η_t yields regret upper bound of

$$\frac{1}{2}D\left(\sqrt{2\sum_{t=1}^{T}\|g_t\|^2} + \sum_{t=1}^{T}\frac{\|g_t\|^2}{\sqrt{2\sum_{i=1}^{t}\|g_i\|^2}}\right).$$

[14] If $\eta_t > \eta_{t+1}$ for some t, then we can further reduce $\mathcal{B}$ by making η_t smaller

Proof

We will show that this is upper bounded by $\sqrt{2}R_{\min}$. For this, we use **Lemma 1**:

Lemma 1

For $x_1, x_2, \ldots x_n \in \mathbb{R}_{\geq 0}$:

$$\sum_{i=1}^{n} \frac{x_i}{\sqrt{\sum_{j=1}^{i} x_j}} \leq 2\sqrt{\sum_{i=1}^{n} x_i}$$

Using **Lemma 1**, we bound the second term:

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{\sqrt{2\sum_{i=1}^{t} \|g_i\|^2}} \leq \sqrt{2\sum_{t=1}^{T} \|g_t\|^2}$$

This gives an improved regret bound of $\sqrt{2}R_{\min}$.

OGD with per-coordinate learning rate

OGD with per-coordinate learning rate Algorithm

- Input: Feasible Set $\mathcal{F}$, T
- Initialize: $x_1 = 0$, $D_i = b_i - a_i$
- for $t = 1$ to T do:
  - Play x_t and observe the cost $f_t(x_t)$
  - Update and Project:

$$y_{t+1,i} = x_{t,i} - \eta_{t,i} g_{t,i}$$
$$x_{t+1} = \Pi_{\mathcal{K}}(y_{t+1})$$

  - Here x_t is a vector whose i^{th} component is $x_{t,i}$ and g_t is a subgradient of f_t at x_t.

OGD Regret Bounds with per-coordinate LRs

Theorem 4

Let $\mathcal{F} = \times_{i=1}^{n}$. Then this Algorithm has regret bounded by $\sum_{i=1}^{n} \mathcal{B}_i\left(\{\eta_{t,i}\}\right)$, where

$$\mathcal{B}_i\left(\{\eta_{t,i}\}\right) := D_i^2 \frac{1}{2\eta_{T,i}} + \frac{1}{2}\sum_{t=1}^{T} g_{t,i}^2 \eta_{t,i} = \sqrt{2} R_{\min,i}$$

Proof

Theorem 4

Let $\mathcal{F} = \times_{i=1}^{n}$. Then this Algorithm has regret bounded by $\mathcal{B}\left(\{\eta_t\}\right) := \sum_{i=1}^{n} \mathcal{B}_i\left(\{\eta_{t,i}\}\right)$, where

$$\mathcal{B}_i\left(\{\eta_{t,i}\}\right) := D_i^2 \frac{1}{2\eta_{T,i}} + \frac{1}{2}\sum_{t=1}^{T} g_{t,i}^2 \eta_{t,i} = \sqrt{2} R_{\min,i}$$

Proof: Since this algorithm only uses subgradient g_t every iteration, we can assume WLOG that f_t is linear and instead use $f_t^L(x) = g_t.x$, to find a regret upper bound[15].

- Since $\mathcal{F}$ is a hypercube, the projector operator independently projects onto $D_i = [a_i, b_i]$.
- This special case is same as solving a separate OCO problem per coordinate i, where at t^{th} iteration, the loss function is f_t^L.

[15]This is because $\text{Regret}_T \leq \text{Regret}_T^L$, by convexity

Proof (cont.)

This means for each i:

- Set η_t such that $\eta_{t,i} = \frac{D_i}{\sqrt{2\sum_{s=1}^{t} g_{s,i}^2}}$ and by using **Theorem 3**, we have the following bound [16]:

$$\text{Regret}_{T,i}^{L} := \sum_{t=1}^{T} g_{t,i} x_{t,i} - \min_{y \in D_i} \{ \sum_{t=1}^{T} g_{t,i} y_i \}$$

$$\text{Regret}_{T,i}^{L} \leq \mathcal{B}_i\left(\{\eta_{t,i}\}\right) = D_i^2 \frac{1}{2\eta_{T,i}} + \frac{1}{2} \sum_{t=1}^{T} g_{t,i}^2 \eta_{t,i}$$

$$\mathcal{B}_i\left(\{\eta_{t,i}\}\right) \leq D_i \sqrt{2 \sum_{t=1}^{T} g_{t,i}^2} = \sqrt{2} R_{\min,i}$$

[16] $R_{\min,i} = \min_{\{\eta_{t,i}\}} \{\mathcal{B}_i\left(\{\eta_{t,i}\}\right)\}$

Proof (cont.)

- Since $\mathcal{F}$ is a hypercube, this means we collect the combined regret as:

$$\text{Regret}_T \leq \text{Regret}_T^L = \sum_{i=1}^{n} \text{Regret}_{T,i}^L$$

$$\text{Regret}_T \leq \sum_{i=1}^{n} \mathcal{B}_i\left(\{\eta_{t,i}\}\right) = \sqrt{2}\sum_{i=1}^{n} R_{\min,i}$$

A tighter Regret Bound

Theorem 5

The bounds obtained in **Theorem 4** is a tighter bound than that obtained in **Theorem 3**, i.e. we show that[a] [b]

$$\sum_{i=1}^{n} D_i \sqrt{2 \sum_{t=1}^{T} g_{t,i}^2} \leq D \sqrt{2 \sum_{t=1}^{T} \|g_t\|^2}$$

where $D = \sqrt{\sum_{i=1}^{n} D_i^2}$ is the diameter of $\mathcal{F}$.

[a] LHS is the bound obtained by using per-coordinate LR
[b] RHS is the improved bound obtained using global LR

Proof

Theorem 5

$$\sum_{i=1}^{n} D_i \sqrt{2\sum_{t=1}^{T} g_{t,i}^2} \leq D\sqrt{2\sum_{t=1}^{T} \|g_t\|^2}$$

Proof: Consider vectors $\vec{D} := \{D_i\}_{i=1}^{n}$ and $\vec{G} := \{\sqrt{2\sum_{t=1}^{T} g_{t,i}^2}\}_{i=1}^{n}$.

- LHS simplifies to $\vec{D}.\vec{G}$
- RHS simplifies to $\|\vec{D}\|\|\vec{G}\|$
- By **Cauchy-Schwarz Inequality** $\vec{D}.\vec{G} \leq \|\vec{D}\|\|\vec{G}\|$

Experimental Evalutation

Online Binary Classification using 2 recent algorithms for text classification: **Passive Aggressive** algorithm (PA) and **Confidence Weighted** algorithm (CW). Here are the results:

Data	Global	Per-Coord	CW	PA
Hinge loss				
BOOKS	0.606	**0.545**	0.871	0.672
DVD	0.576	**0.529**	0.851	0.637
ELECTRONICS	0.509	**0.452**	0.802	0.555
KITCHEN	0.470	**0.419**	0.787	0.520
NEWS	0.171	**0.140**	0.512	0.245
RCV1	0.076	**0.070**	0.542	0.094
Fraction of mistakes				
BOOKS	0.259	**0.211**	0.215	0.254
DVD	0.238	0.208	**0.203**	0.240
ELECTRONICS	0.209	**0.175**	0.177	0.194
KITCHEN	0.180	**0.151**	0.153	0.175
NEWS	0.064	**0.050**	0.054	0.060
RCV1	0.027	**0.025**	0.039	0.034

Further Exploration

- For α-strongly convex functions we have a regret bound of $\frac{G}{2\alpha}(1 + \log T)$.
- Improved global regret bounds on α-strongly convex functions.
- Improved per-coordinate regret bounds on α-strongly convex functions.

Thank You