Partially Lazy Gradient Descent for Smoothed Online Learning

Naram Mhaisen and George Iosifidis

TU Delft Delft University of Technology

Faculty of Electrical Engineering, Mathematics, and Computer Science.

Department of Software Technology, Networked Systems Group.

1 Problem setup

Smoothed Online Convex Optimization :

For T slots:

- Execute $\boldsymbol{x}_t$
- Incur $f_t(\boldsymbol{x}_t) + \lambda\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|$
- Receive $\boldsymbol{g}_t \in \partial f_t(\boldsymbol{x}_t)$

Aim: minimize $\mathcal{R}_T = \sum_{t=1}^{T}(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{u}_t) + \lambda\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|)$

Fact: $\mathcal{R}_T = \Omega\left(\sqrt{T(P_T{+}1)}\right), \quad P_T = \sum_{t=1}^{T-1}\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|$

2 GD: Two variants

GD: $\boldsymbol{x}_{t+1} = \Pi(\boldsymbol{x}_t - \eta\,\boldsymbol{g}_t)$, $\boldsymbol{x}_t = \Pi(\boldsymbol{x}_{t-1} - \eta\,\boldsymbol{g}_{t-1})$

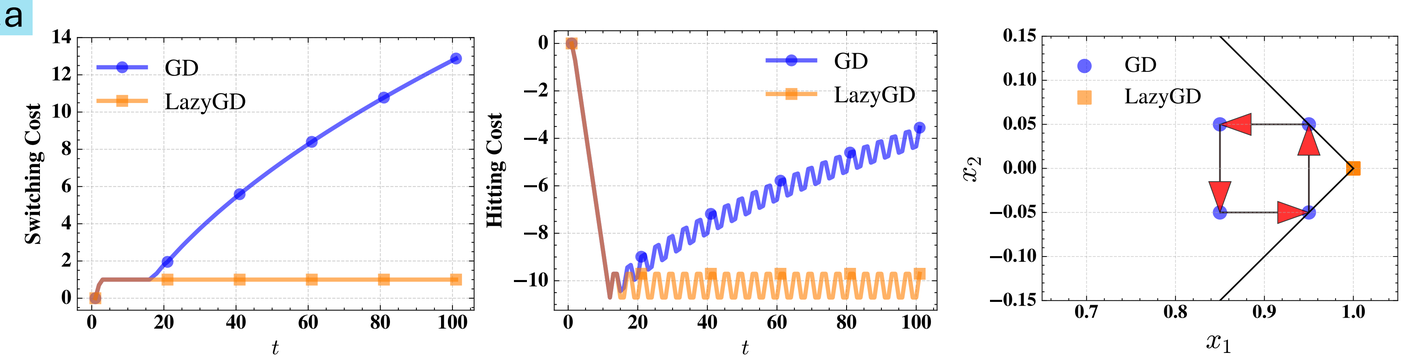
LazyGD: $\boldsymbol{x}_{t+1} = \Pi(-\eta\,\boldsymbol{g}_{1:t})$, $\boldsymbol{x}_t = \Pi(-\eta\,\boldsymbol{g}_{1:t-1})$

Both ensure : $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = \mathcal{O}(\eta)$

i.e., can be made sublinear.

3 Practical insight: Different switching behavior

3.a



Example: $g_t = (-1, 0)$ for $t = 1, \ldots, 12$ (to reach the boundary), then cycles clockwise through $(1,0) \to (0,1) \to (-1,0) \to (0,-1)$.

Reason:

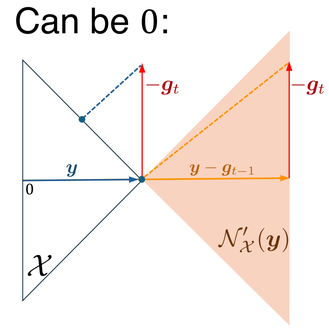
LazyGD: $\|\Pi(-\eta\boldsymbol{g}_{1:t}) - \Pi(-\eta\boldsymbol{g}_{1:t-1})\|$ — Small change in direction

GD: $\|\Pi(\boldsymbol{x}_t - \eta\boldsymbol{g}_t) - \Pi(\boldsymbol{x}_t)\|$ — Large change in direction

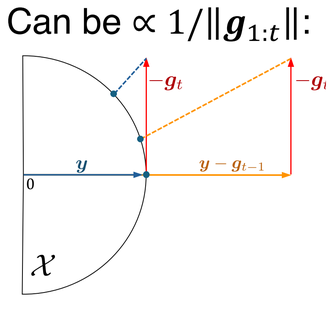
3.b

When $\hat{\boldsymbol{a}} \approx \hat{\boldsymbol{b}}$, $\|\Pi(\boldsymbol{a}) - \Pi(\boldsymbol{b})\|$

Can be 0:



Can be $\propto 1/\|\boldsymbol{g}_{1:t}\|$:



- Preference for laziness (aggregation).
- Obstacle: $\boldsymbol{x}_{t+1} = \Pi(\boldsymbol{x}_0 - \eta\boldsymbol{g}_{1:t}) \Rightarrow \mathcal{R}_T = \Omega(T)$

What is the maximum possible laziness (aggregation) while maintaining $\mathcal{R}_T = \mathcal{O}\left(\sqrt{T(P_T{+}1)}\right)$?

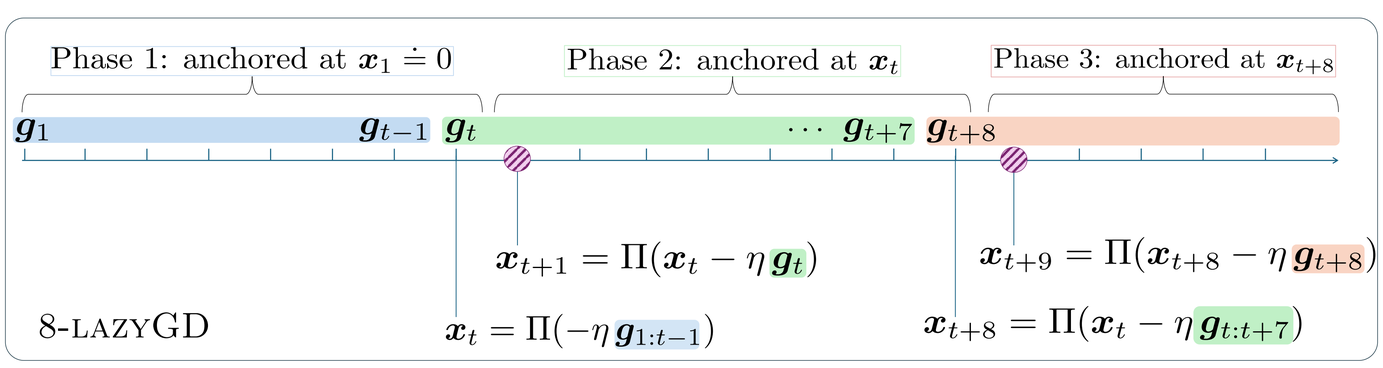
4 Proposal: Unified view

4.a

$$\boldsymbol{x}_{t+1} = \Pi\big(\underbrace{\boldsymbol{x}_{t-n_t} - \eta\boldsymbol{g}_{t-n_t:t}}_{\bar{\boldsymbol{y}}_{t+1}}\big), \quad n_t = (t-1) \mod (k),$$

- $k = 1 \Rightarrow n_t = 0, \forall t \Rightarrow \quad \bar{\boldsymbol{y}}_{t+1} = \boldsymbol{x}_t - \eta\boldsymbol{g}_t$ (GD)
- $k = T \Rightarrow n_t = t{-}1, \forall t \Rightarrow \quad \bar{\boldsymbol{y}}_{t+1} = -\eta\boldsymbol{g}_{1:t}$ (LazyGD)
- In general:

$$[T] = \{\overbrace{1, \ldots, k}^{[p_1]}, \overbrace{k{+}1, \ldots, 2k}^{[p_2]}, \overbrace{2k{+}1, \ldots, 3k}^{[p_3]}, \ldots\}.$$ Within each $[p_i]$: $\boldsymbol{x}_{t+1} = \Pi(\boldsymbol{x}_{p_i} - \eta\boldsymbol{g}_{p_i:t})$.



4.b

- For any k: $\mathcal{R}_T = \Omega\left(\max\left(\sqrt{T(P_T+1)},\ k\,P_T\right)\right)$.
- $k = \Theta(\sqrt{T/P_T})$ is possible.
- Setting $k = \Theta(\sqrt{T/P_T})$ and $\eta = \Theta(\sqrt{P_T/T})$, $\mathcal{R}_T = \mathcal{O}(\sqrt{(P_T+1)T})$.
- With the lazy switching features:

$\bar{\boldsymbol{y}}_{t+1} \in \mathcal{N}'_{\mathcal{X}}(\boldsymbol{x}_t) \Rightarrow \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = 0$

$\|\bar{\boldsymbol{y}}_{t+1}\| \geq R + \frac{\alpha G\, n_t}{\sigma} \Rightarrow \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = \mathcal{O}(\frac{1}{k}\log(1 + \frac{k}{\sigma}))$

(more likely when $\bar{\boldsymbol{y}}_{t+1}$ is an aggregate)

5 Tuning & numerical examples

SAder-k: SAder with k-LazyGD base learners:

$$\boldsymbol{x}_{t+1} = \sum_i w_{t+1}\,\boldsymbol{x}^{(i)}_{t+1}$$

$\boldsymbol{x}^{(N)}_{t+1} = \Pi(x^{(N)}_t - \eta^{(N)}\boldsymbol{g}_t)$

$\vdots$

$\boldsymbol{x}^{(2)}_{t+1} = \Pi(x^{(2)}_{t-\sqrt{T/4}} - \eta^{(2)}\boldsymbol{g}_{t-\sqrt{T/4}:t})$

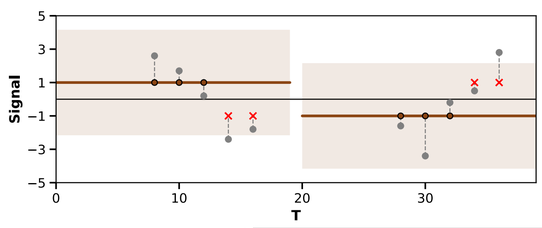
$\boldsymbol{x}^{(1)}_{t+1} = \Pi(x^{(1)}_{t-\sqrt{T/2}} - \eta^{(1)}\boldsymbol{g}_{t-\sqrt{T/2}:t})$

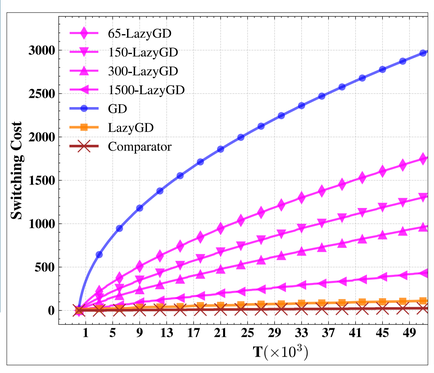
$\boldsymbol{x}^{(0)}_{t+1} = \Pi(x^{(0)}_{t-\sqrt{T}} - \eta^{(0)}\boldsymbol{g}_{t-\sqrt{T}:t})$

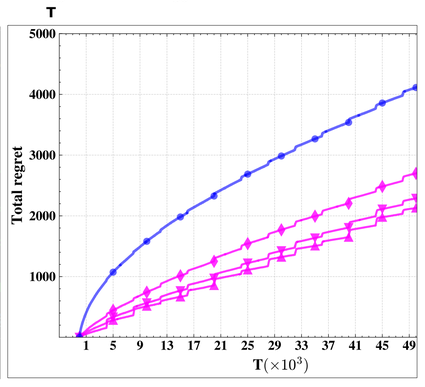
$\boldsymbol{x}^{(L)}_{t+1} = \Pi(x^{(L)}_0 - \eta^{(L)}\boldsymbol{g}_{1:t})$

Hedging on a set of learning rates and aggregation scales.

Adversarial mean switching + Stochastic noise







Adversarial switching + Outliers

