Adaptive Regret for Online Learning

Lijun Zhang

Dept. of Computer Science and Technology, Nanjing University

Microsoft Research Asia Academic Day 2019

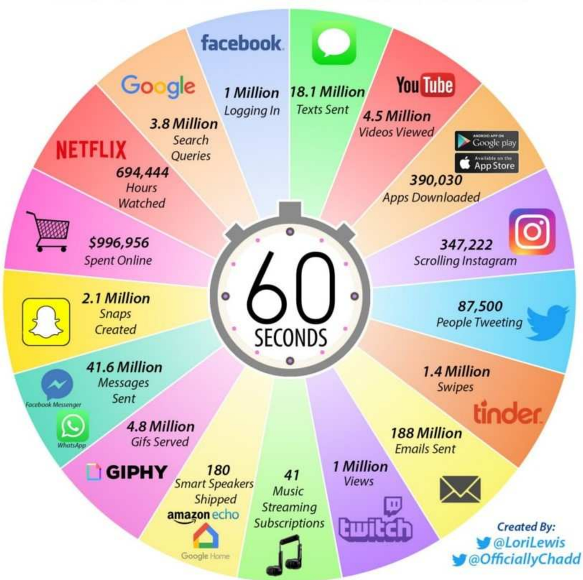
Outline

LAMDA Learning And Mining from DatA

Outline

1. Introduction
2. Learning in Dynamic Environments
   - Adaptive Regret
   - Adaptive Algorithms
3. Our Contributions
   - Faster Rates for Adaptive Regret
   - Universal Algorithm for Adaptive Regret
4. Conclusion

What Happens in an Internet Minute?



https://www.garyfox.co/what-happens-internet-minute-2019/

Online Learning

- Online Convex Optimization [Zinkevich, 2003]

1: **for** $t = 1, 2, \ldots, T$ **do**

4: **end for**

Online Learning

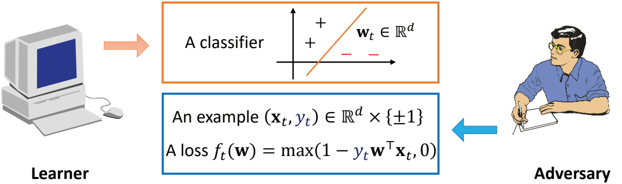
- Online Convex Optimization [Zinkevich, 2003]

for $t = 1, 2, \ldots, T$ **do**

Learner picks a decision $\mathbf{w}_t \in \mathcal{W}$

Adversary chooses a convex function $f_t(\cdot) : \mathcal{W} \mapsto \mathbb{R}$

end for



Online Learning

- Online Convex Optimization [Zinkevich, 2003]

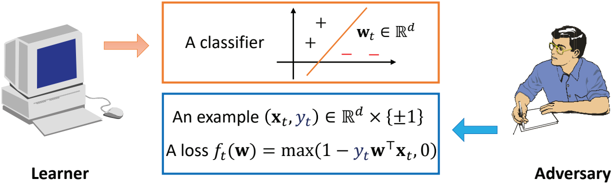
for $t = 1, 2, \ldots, T$ **do**

  Learner picks a decision $\mathbf{w}_t \in \mathcal{W}$

  Adversary chooses a convex function $f_t(\cdot) : \mathcal{W} \mapsto \mathbb{R}$

  Learner suffers loss $f_t(\mathbf{w}_t)$ and updates $\mathbf{w}_t$

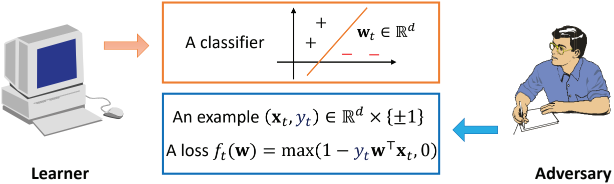
end for



Online Learning

- Online Convex Optimization [Zinkevich, 2003]

for $t = 1, 2, \ldots, T$ **do**
    Learner picks a decision $\mathbf{w}_t \in \mathcal{W}$
    Adversary chooses a convex function $f_t(\cdot) : \mathcal{W} \mapsto \mathbb{R}$
    Learner suffers loss $f_t(\mathbf{w}_t)$ and updates $\mathbf{w}_t$
end for



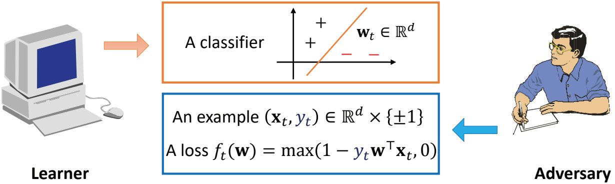
- Cumulative Loss

$$\text{Cumulative Loss} = \sum_{t=1}^{T} f_t(\mathbf{w}_t)$$

Online Learning

- Online Convex Optimization [Zinkevich, 2003]

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner picks a decision $\mathbf{w}_t \in \mathcal{W}$
Adversary chooses a convex function $f_t(\cdot) : \mathcal{W} \mapsto \mathbb{R}$
3: Learner suffers loss $f_t(\mathbf{w}_t)$ and updates $\mathbf{w}_t$
4: **end for**



- Cumulative Loss

$$\text{Regret} = \underbrace{\sum_{t=1}^{T} f_t(\mathbf{w}_t)}_{\text{Cumulative Loss of Online Learner}} - \underbrace{\min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w})}_{\text{Minimal Loss of Offline Learner}}$$

Online Gradient Descent (OGD)

- Algorithm

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner picks a decision $\mathbf{w}_t \in \mathcal{W}$
Adversary chooses a convex function $f_t(\cdot) : \mathcal{W} \mapsto \mathbb{R}$
3: Learner suffers loss $f_t(\mathbf{w}_t)$ and

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}} \left[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t) \right]$$

4: **end for**

- The Projection Operator

$$\Pi_{\mathcal{W}}[\mathbf{x}] = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \|\mathbf{w} - \mathbf{x}\|_2$$

LAMDA Learning And Mining from DatA

Online Convex Optimization

- Convex Functions [Zinkevich, 2003]
  - Online Gradient Descent with $\eta_t = 1/\sqrt{t}$

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = O\left(\sqrt{T}\right)$$

- Strongly Convex Functions [Hazan et al., 2007]
  - Online Gradient Descent with $\eta_t = 1/(\lambda t)$

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = O\left(\log T\right)$$

- Exponentially Concave Functions [Hazan et al., 2007]
  - Online Newton Step

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = O\left(d \log T\right)$$

Outline

The Challenge

Regret → Static Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{w}_*)$$

where $\mathbf{w}_* \in \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w})$

- One of the decision is reasonably good during T rounds

LAMDA Learning And Mining from DatA

The Challenge

Regret → Static Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{w}_*)$$

where $\mathbf{w}_* \in \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w})$

- One of the decision is reasonably good during T rounds

Dynamic Environments

Different decisions will be good in different periods

- Recommendation: the interests of a user could change
- Stock market: the best stock changes over time

Outline

Adaptive Regret

Adaptive Regret

[Hazan and Seshadhri, 2007, Daniely et al., 2015]

$$R(T,\tau) = \max_{[s,s+\tau-1]\subseteq[T]} \left(\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}) \right)$$

- Minimize the static regret over all intervals of length τ

LAMDA Learning And Mining from DatA

Adaptive Regret

Adaptive Regret

[Hazan and Seshadhri, 2007, Daniely et al., 2015]

$$R(T,\tau) = \max_{[s,s+\tau-1]\subseteq[T]} \left(\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}) \right)$$

- Minimize the static regret over all intervals of length τ

$$f_1(\cdot),\, f_2(\cdot), \ldots, f_\tau(\cdot), f_{\tau+1}(\cdot), \ldots, f_s(\cdot),\, f_{s+1}(\cdot), \ldots, f_{s+\tau-1}(\cdot), f_{s+\tau}(\cdot), \ldots$$

Adaptive Regret

Adaptive Regret

[Hazan and Seshadhri, 2007, Daniely et al., 2015]

$$R(T,\tau) = \max_{[s,s+\tau-1]\subseteq[T]} \left(\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}) \right)$$

- Minimize the static regret over all intervals of length τ

$$\underbrace{\boxed{f_1(\cdot),\, f_2(\cdot), \ldots, f_\tau(\cdot)}}_{\sum_{t=1}^{\tau} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{\tau} f_t(\mathbf{w})}, f_{\tau+1}(\cdot), \ldots, f_s(\cdot),\, f_{s+1}(\cdot), \ldots, f_{s+\tau-1}(\cdot), f_{s+\tau}(\cdot), \ldots$$

LAMDA Learning And Mining from DatA

Adaptive Regret

Adaptive Regret

[Hazan and Seshadhri, 2007, Daniely et al., 2015]

$$R(T,\tau) = \max_{[s,s+\tau-1]\subseteq[T]} \left(\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}) \right)$$

- Minimize the static regret over all intervals of length τ

$$\overbrace{\quad}^{\sum_{t=2}^{\tau+1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=2}^{\tau+1} f_t(\mathbf{w})}$$

$$\underbrace{f_1(\cdot), f_2(\cdot), \ldots, f_\tau(\cdot)}_{\sum_{t=1}^{\tau} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{\tau} f_t(\mathbf{w})}, f_{\tau+1}(\cdot), \ldots, f_s(\cdot), f_{s+1}(\cdot), \ldots, f_{s+\tau-1}(\cdot), f_{s+\tau}(\cdot), \ldots$$

Adaptive Regret

Adaptive Regret

[Hazan and Seshadhri, 2007, Daniely et al., 2015]

$$R(T,\tau) = \max_{[s,s+\tau-1]\subseteq[T]} \left(\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}) \right)$$

- Minimize the static regret over all intervals of length τ

$$\overbrace{\sum_{t=2}^{\tau+1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=2}^{\tau+1} f_t(\mathbf{w})}$$

$$\underbrace{f_1(\cdot), f_2(\cdot), \ldots, f_\tau(\cdot)}_{\sum_{t=1}^{\tau} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{\tau} f_t(\mathbf{w})}, f_{\tau+1}(\cdot), \ldots, \underbrace{\boxed{f_s(\cdot), f_{s+1}(\cdot), \ldots, f_{s+\tau-1}(\cdot)}}_{\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w})}, f_{s+\tau}(\cdot), \ldots$$

Adaptive Regret

Adaptive Regret

[Hazan and Seshadhri, 2007, Daniely et al., 2015]

$$R(T,\tau) = \max_{[s,s+\tau-1]\subseteq[T]} \left(\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}) \right)$$

- Minimize the static regret over all intervals of length τ

$$f_1(\cdot), \underbrace{\overbrace{f_2(\cdot), \ldots, f_\tau(\cdot)}^{\sum_{t=2}^{\tau+1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=2}^{\tau+1} f_t(\mathbf{w})}}_{\sum_{t=1}^{\tau} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{\tau} f_t(\mathbf{w})}, f_{\tau+1}(\cdot), \ldots, \underbrace{f_s(\cdot), \overbrace{\boxed{f_{s+1}(\cdot), \ldots, f_{s+\tau-1}(\cdot)}, f_{s+\tau}(\cdot)}^{\sum_{t=s+1}^{s+\tau} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s+1}^{s+\tau} f_t(\mathbf{w})}}_{\sum_{t=s}^{s+\tau-1} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=s}^{s+\tau-1} f_t(\mathbf{w})}, \ldots$$

Outline

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t		1		2		3		4		5		6		7		$\cdots$
$\mathcal{I}_0$	[		] [		] [		] [		] [		] [		] [		]	$\cdots$
$\mathcal{I}_1$			[				] [				] [				]	$\cdots$
$\mathcal{I}_2$							[								]	$\cdots$

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	$\cdots$
$\mathcal{I}_0$	[OGD(f_1)]							$\cdots$
$\mathcal{I}_1$								$\cdots$
$\mathcal{I}_2$								$\cdots$

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	...
$\mathcal{I}_0$		[OGD(f_2)]						...
$\mathcal{I}_1$		[OGD(f_2,		)]				...
$\mathcal{I}_2$								...

The rows $\mathcal{I}_0$ and $\mathcal{I}_1$ are joined by a brace labelled "meta-algorithm".

- A Meta-algorithm
  - Strongly Adaptive Online Learner [Daniely et al., 2015]
  - Track the best expert on the fly

LAMDA Learning And Mining from DatA

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	$\cdots$
$\mathcal{I}_0$			[OGD(f_3)]	meta-algorithm				$\cdots$
$\mathcal{I}_1$		[OGD(f_2,	f_3)]					$\cdots$
$\mathcal{I}_2$								$\cdots$

- A Meta-algorithm
  - Strongly Adaptive Online Learner [Daniely et al., 2015]
  - Track the best expert on the fly

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	$\cdots$
$\mathcal{I}_0$			meta-algorithm	[OGD(f_4)]				$\cdots$
$\mathcal{I}_1$				[OGD(f_4,	)]			$\cdots$
$\mathcal{I}_2$				[OGD(f_4,			)]	$\cdots$

- A Meta-algorithm
  - Strongly Adaptive Online Learner [Daniely et al., 2015]
  - Track the best expert on the fly

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	$\cdots$
$\mathcal{I}_0$			meta-algorithm $\{$		[OGD(f_5)]			$\cdots$
$\mathcal{I}_1$				[OGD(f_4,	f_5)]			$\cdots$
$\mathcal{I}_2$				[OGD(f_4,	f_5,		)]	$\cdots$

- A Meta-algorithm
  - Strongly Adaptive Online Learner [Daniely et al., 2015]
  - Track the best expert on the fly

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	$\cdots$
$\mathcal{I}_0$						[OGD(f_6)]		$\cdots$
$\mathcal{I}_1$			meta-algorithm			[OGD(f_6,	)]	$\cdots$
$\mathcal{I}_2$				[OGD(f_4,	f_5,	f_6,	)]	$\cdots$

- A Meta-algorithm
  - Strongly Adaptive Online Learner [Daniely et al., 2015]
  - Track the best expert on the fly

LAMDA Learning And Mining from DatA

The General Framework

- An Expert-algorithm
  - Online Gradient Descent (OGD) [Hazan et al., 2007]

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t)]$$

- A Set of Intervals
  - Geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	$\cdots$
$\mathcal{I}_0$							$[\text{OGD}(f_7)]$	$\cdots$
$\mathcal{I}_1$			meta-algorithm			$[\text{OGD}(f_6,$	$f_7)]$	$\cdots$
$\mathcal{I}_2$				$[\text{OGD}(f_4,$	$f_5,$	$f_6,$	$f_7)]$	$\cdots$

  - Number of experts in round t: $O(\log t)$

- A Meta-algorithm
  - Strongly Adaptive Online Learner [Daniely et al., 2015]
  - Track the best expert on the fly

Theoretical Guarantees

- Convex Functions [Jun et al., 2017]

$$R(T, \tau) = O\left(\sqrt{\tau \log T}\right)$$

- Strongly Convex Functions [Zhang et al., 2018]

$$R(T, \tau) = O\left(\log \tau \log T\right)$$

- Exponentially Concave Functions [Hazan and Seshadhri, 2007]

$$R(T, \tau) = O\left(d \log \tau \log T\right)$$

Outline

Our Work on Adaptive Regret

1. **Lijun Zhang**, Guanghui Wang, Wei-Wei Tu, and Zhi-Hua Zhou. Dual Adaptivity: A Universal Algorithm for Minimizing the Adaptive Regret of Convex Functions. In *https://arxiv.org/abs/1906.10851*, 2019.
2. **Lijun Zhang**, Tie-Yan Liu, and Zhi-Hua Zhou. Adaptive Regret of Convex and Smooth Functions. In *Proceedings of the 36th International Conference on Machine Learning* (ICML), 2019.
3. **Lijun Zhang**, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Dynamic Regret of Strongly Adaptive Methods. In *Proceedings of the 35th International Conference on Machine Learning* (ICML), 2018.
4. Guanghui Wang, Dakuan Zhao, and **Lijun Zhang**. Minimizing Adaptive Regret with One Gradient per Iteration. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (IJCAI), 2018.

LAMDA Learning And Mining from DatA

Existing Results on Adaptive Regret

- Convex Functions [Jun et al., 2017]

$$R(T,\tau) = O\left(\sqrt{\tau \log T}\right)$$

- Strongly Convex Functions [Zhang et al., 2018]

$$R(T,\tau) = O\left(\log \tau \log T\right)$$

- Exponentially Concave Functions [Hazan and Seshadhri, 2007]

$$R(T,\tau) = O\left(d \log \tau \log T\right)$$

Existing Results on Adaptive Regret

- Convex Functions [Jun et al., 2017]

$$R(T,\tau) = O\left(\sqrt{\tau \log T}\right)$$

- Strongly Convex Functions [Zhang et al., 2018]

$$R(T,\tau) = O\left(\log \tau \log T\right)$$

- Exponentially Concave Functions [Hazan and Seshadhri, 2007]

$$R(T,\tau) = O\left(d \log \tau \log T\right)$$

Limitations

1. The regret bound is problem-independent
2. Existing algorithms are not universal

Outline

Motivations

- Convex Functions [Zinkevich, 2003]
  - Online Gradient Descent with $\eta_t = 1/\sqrt{t}$

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = O\left(\sqrt{T}\right)$$

Motivations

- Convex Functions [Zinkevich, 2003]
  - Online Gradient Descent with $\eta_t = 1/\sqrt{t}$

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = O\left(\sqrt{T}\right)$$

- Convex and Smooth Functions [Srebro et al., 2010]
  - Online Gradient Descent with prior knowledge

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{w}) = O\left(\sqrt{\sum_{t=1}^{T} f_t(\mathbf{w})}\right) = O\left(\sqrt{T}\right)$$

  - The bound could be tighter when $\sqrt{\sum_{t=1}^{T} f_t(\mathbf{w})}$ is small
  - The regret bound is problem-dependent

The Problem

Question

Can smoothness be exploited to boost the adaptive regret?

The Problem

Question

Can smoothness be exploited to boost the adaptive regret?

- Regret over an interval $[r, s]$

$$\text{Regret}\left([r, s]\right) = \sum_{t=r}^{s} f_t(\mathbf{w}_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=r}^{s} f_t(\mathbf{w})$$

- Convex Functions [Jun et al., 2017]

$$\text{Regret}\left([r, s]\right) = O\left(\sqrt{(s-r)\log s}\right)$$

$$\Rightarrow R(T, \tau) = O\left(\sqrt{\tau \log T}\right)$$

Our Results [Zhang et al., 2019a]

- Convex and Smooth Functions

$$\text{Regret}\left([r,s]\right) = O\left(\sqrt{\left(\sum_{t=r}^{s} f_t(\mathbf{w})\right) \log s \cdot \log(s-r)}\right)$$

  - Become tighter when $\sum_{t=r}^{s} f_t(\mathbf{w})$ is small

- Convex Functions [Jun et al., 2017]

$$\text{Regret}\left([r,s]\right) = O\left(\sqrt{(s-r)\log s}\right)$$

Our Results [Zhang et al., 2019a]

- Convex and Smooth Functions

$$\text{Regret}\left([r,s]\right) = O\left(\sqrt{\left(\sum_{t=r}^{s} f_t(\mathbf{w})\right) \log s \cdot \log(s-r)}\right)$$

  - Become tighter when $\sum_{t=r}^{s} f_t(\mathbf{w})$ is small

- Convex Functions [Jun et al., 2017]

$$\text{Regret}\left([r,s]\right) = O\left(\sqrt{(s-r)\log s}\right)$$

- Convex and Smooth Functions

$$\text{Regret}\left([r,s]\right) = O\left(\sqrt{\left(\sum_{t=r}^{s} f_t(\mathbf{w})\right) \log \sum_{t=1}^{s} f_t(\mathbf{w}) \cdot \log \sum_{t=r}^{s} f_t(\mathbf{w})}\right)$$

  - Fully problem-dependent

The Algorithm

- An Expert-algorithm
  - Scale-free online gradient descent [Orabona and Pál, 2018]
  - Can exploit smoothness automatically
- A Set of Intervals
  - Compact geometric covering intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	⋯
$\mathcal{C}_0$	[ ]		[ ]		[ ]		[ ]		[ ]		[ ]		[ ]		[ ]		[ ]		⋯
$\mathcal{C}_1$		[	]			[	]			[	]			[	]			[	⋯
$\mathcal{C}_2$				[			]					[			]				⋯
$\mathcal{C}_3$								[							]				⋯
$\mathcal{C}_4$																[			⋯

- A Meta-algorithm
  - AdaNormalHedge [Luo and Schapire, 2015]
  - Attain a small-loss regret and support sleeping experts

Outline

1. Introduction
2. Learning in Dynamic Environments
   - Adaptive Regret
   - Adaptive Algorithms
3. Our Contributions
   - Faster Rates for Adaptive Regret
   - Universal Algorithm for Adaptive Regret
4. Conclusion

Our Results [Zhang et al., 2019b]

Second-Order Adaptive Regret Bounds

$$\text{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

- Adapt to different types of loss functions

Our Results [Zhang et al., 2019b]

- Second-Order Adaptive Regret Bounds

$$\text{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

  - Adapt to different types of loss functions
- For General Convex Functions

$$\text{Regret}\left([r,s]\right) = O\left(\sqrt{(s-r)\log T}\right)$$

Our Results [Zhang et al., 2019b]

- Second-Order Adaptive Regret Bounds

$$\text{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

  - Adapt to different types of loss functions
- For α-Exp-Concave Functions

$$\text{Regret}\left([r,s]\right) = O\left((d \log(s-r) \log T)/\alpha\right)$$

  - Agnostic to α

Our Results [Zhang et al., 2019b]

- Second-Order Adaptive Regret Bounds

$$\text{Regret}([r,s]) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

  - Adapt to different types of loss functions
- For λ-Strongly Convex Functions

$$\text{Regret}([r,s]) = O((\log(s-r)\log T)/\lambda)$$

  - Agnostic to λ

Our Results [Zhang et al., 2019b]

Second-Order Adaptive Regret Bounds

$$\text{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

- Adapt to different types of loss functions
- Allow the type of loss functions to switch between rounds

Our Results [Zhang et al., 2019b]

- Second-Order Adaptive Regret Bounds

$$\text{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

- Adapt to different types of loss functions
- Allow the type of loss functions to switch between rounds

$$\ldots, f_{r_1}(\cdot), \ldots, f_{s_1}(\cdot), \ldots, f_{r_2}(\cdot), \ldots, f_{s_2}(\cdot), \ldots, f_{r_3}(\cdot), \ldots, f_{s_3}(\cdot), \ldots$$

Our Results [Zhang et al., 2019b]

Second-Order Adaptive Regret Bounds

$$\text{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

- Adapt to different types of loss functions
- Allow the type of loss functions to switch between rounds

$$\ldots, \underbrace{f_{r_1}(\cdot), \ldots, f_{s_1}(\cdot)}_{\text{convex}}, \ldots, f_{r_2}(\cdot), \ldots, f_{s_2}(\cdot), \ldots, f_{r_3}(\cdot), \ldots, f_{s_3}(\cdot), \ldots$$

$O(\sqrt{(s_1 - r_1)\log(T)})$

Our Results [Zhang et al., 2019b]

Second-Order Adaptive Regret Bounds

$$\mathrm{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

- Adapt to different types of loss functions
- Allow the type of loss functions to switch between rounds

$$\ldots, \underbrace{f_{r_1}(\cdot), \ldots, f_{s_1}(\cdot)}_{\substack{\text{convex} \\ O(\sqrt{(s_1 - r_1)\log(T)})}}, \ldots, \underbrace{f_{r_2}(\cdot), \ldots, f_{s_2}(\cdot)}_{\substack{\text{strongly convex} \\ O(\log(s_2 - r_2)\log T)}}, \ldots, f_{r_3}(\cdot), \ldots, f_{s_3}(\cdot), \ldots$$

LAMDA Learning And Mining from DatA

Our Results [Zhang et al., 2019b]

- Second-Order Adaptive Regret Bounds

$$\text{Regret}\left([r,s]\right) \leq \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

$$= \begin{cases} \widetilde{O}\left(\sqrt{d \sum_{t=r}^{s} \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle^2}\right) \\ \widetilde{O}\left(\sqrt{\sum_{t=r}^{s} \|\mathbf{w}_t - \mathbf{w}\|^2}\right) \end{cases}$$

- Adapt to different types of loss functions
- Allow the type of loss functions to switch between rounds

$$\ldots, \underbrace{f_{r_1}(\cdot), \ldots, f_{s_1}(\cdot)}_{\substack{\text{convex} \\ O(\sqrt{(s_1-r_1)\log(T)})}}, \ldots, \underbrace{f_{r_2}(\cdot), \ldots, f_{s_2}(\cdot)}_{\substack{\text{strongly convex} \\ O(\log(s_2-r_2)\log T)}}, \ldots, \underbrace{\boxed{f_{r_3}(\cdot), \ldots, f_{s_3}(\cdot)}}_{\substack{\text{exp-concave} \\ O(d\log(s_3-r_3)\log T)}}, \ldots$$

The Algorithm

- Two Types of Experts
  - $E_{\mathcal{J}}^{\eta}$: ONS for exp-concave surrogate loss:

$$\ell_t^{\eta}(\mathbf{w}) = -\eta\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle^2$$

  - $\widehat{E}_{\mathcal{J}}^{\eta}$: OGD for strongly convex surrogate loss:

$$\hat{\ell}_t^{\eta}(\mathbf{w}) = -\eta\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2 G^2\|\mathbf{w}_t - \mathbf{w}\|^2$$

The Algorithm

- Two Types of Experts
  - $E_{\mathcal{J}}^{\eta}$: ONS for exp-concave surrogate loss:

$$\ell_t^{\eta}(\mathbf{w}) = -\eta\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle^2$$

  - $\widehat{E}_{\mathcal{J}}^{\eta}$: OGD for strongly convex surrogate loss:

$$\hat{\ell}_t^{\eta}(\mathbf{w}) = -\eta\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2 G^2\|\mathbf{w}_t - \mathbf{w}\|^2$$

- A Set of Intervals: GC intervals [Daniely et al., 2015]

t 1 2 3 4 5 6 7 8 9 10 11 $\cdots$
$\mathcal{I}_0$ [ ][ ][ ][ ][ ][ ][ ][ ][ ][ ][ ] $\cdots$
$\mathcal{I}_1$ [ ][ ][ ][ ][ ] $\cdots$
$\mathcal{I}_2$ [ ][ ] $\cdots$

The Algorithm

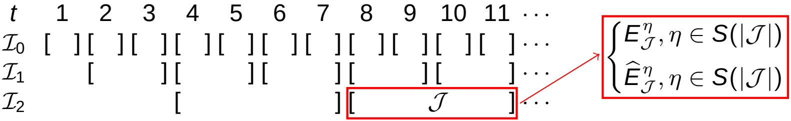
- Two Types of Experts
  - $E_{\mathcal{J}}^{\eta}$: ONS for exp-concave surrogate loss:

$$\ell_t^{\eta}(\mathbf{w}) = -\eta\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2 \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle^2$$

  - $\widehat{E}_{\mathcal{J}}^{\eta}$: OGD for strongly convex surrogate loss:

$$\hat{\ell}_t^{\eta}(\mathbf{w}) = -\eta\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2 G^2 \|\mathbf{w}_t - \mathbf{w}\|^2$$

- A Set of Intervals: GC intervals [Daniely et al., 2015]



$$\mathcal{S}(|\mathcal{J}|) = \left\{ \frac{2^{-i}}{5DG} \;\middle|\; i = 0, 1, \ldots, \left\lceil \frac{1}{2}\log_2 |\mathcal{J}| \right\rceil \right\}$$

The Algorithm

- Two Types of Experts
  - $E_{\mathcal{J}}^{\eta}$: ONS for exp-concave surrogate loss:

$$\ell_t^{\eta}(\mathbf{w}) = -\eta\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2 \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle^2$$

  - $\widehat{E}_{\mathcal{J}}^{\eta}$: OGD for strongly convex surrogate loss:

$$\hat{\ell}_t^{\eta}(\mathbf{w}) = -\eta\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle + \eta^2 G^2 \|\mathbf{w}_t - \mathbf{w}\|^2$$

- A Set of Intervals: GC intervals [Daniely et al., 2015]

t	1	2	3	4	5	6	7	8	9	10	11	$\cdots$
$\mathcal{I}_0$	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	$\cdots$
$\mathcal{I}_1$		[	]	[	]	[	]	[	]	[	]	$\cdots$
$\mathcal{I}_2$				[			]	[	$\mathcal{J}$		]	$\cdots$

$$\begin{cases} E_{\mathcal{J}}^{\eta}, \eta \in S(|\mathcal{J}|) \\ \widehat{E}_{\mathcal{J}}^{\eta}, \eta \in S(|\mathcal{J}|) \end{cases}$$

$$\mathcal{S}(|\mathcal{J}|) = \left\{ \frac{2^{-i}}{5DG} \;\middle|\; i = 0, 1, \ldots, \left\lceil \frac{1}{2}\log_2 |\mathcal{J}| \right\rceil \right\}$$

- A Meta-Algorithm
  - TWEA [van Erven and Koolen, 2016] with sleeping experts

Outline

1. Introduction
2. Learning in Dynamic Environments
   - Adaptive Regret
   - Adaptive Algorithms
3. Our Contributions
   - Faster Rates for Adaptive Regret
   - Universal Algorithm for Adaptive Regret
4. Conclusion

Conclusion and Future Work

- Conclusion
  - Problem-dependent Adaptive Regret
    - Make use of the smoothness automatically
  - Universal Algorithm for Adaptive Regret
    - Convex, exponentially concave, and strongly convex

- Future Work
  - Extension of Previous Studies
    - Problem-dependent & Universal
  - Dynamic Regret

$$R(\mathbf{u}_1, \ldots, \mathbf{u}_T) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t)$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{W}$

Reference I

Thanks!

Daniely, A., Gonen, A., and Shalev-Shwartz, S. (2015).
Strongly adaptive online learning.
In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1405–1411.

Hazan, E., Agarwal, A., and Kale, S. (2007).
Logarithmic regret algorithms for online convex optimization.
Machine Learning, 69(2-3):169–192.

Hazan, E. and Seshadhri, C. (2007).
Adaptive algorithms for online decision problems.
Electronic Colloquium on Computational Complexity, 88.

Jun, K.-S., Orabona, F., Wright, S., and Willett, R. (2017).
Improved strongly adaptive online learning using coin betting.
In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 943–951.

Luo, H. and Schapire, R. E. (2015).
Achieving all with no parameters: Adanormalhedge.
In *Proceedings of The 28th Conference on Learning Theory*, pages 1286–1304.

Orabona, F. and Pál, D. (2018).
Scale-free online learning.
Theoretical Computer Science, 716:50–69.

LAMDA Learning And Mining from DatA

Reference II

Srebro, N., Sridharan, K., and Tewari, A. (2010).
Smoothness, low-noise and fast rates.
In *Advances in Neural Information Processing Systems 23*, pages 2199–2207.

van Erven, T. and Koolen, W. M. (2016).
Metagrad: Multiple learning rates in online learning.
In *Advances in Neural Information Processing Systems 29*, pages 3666–3674.

Zhang, L., Liu, T.-Y., and Zhou, Z.-H. (2019a).
Adaptive regret of convex and smooth functions.
In *Proceedings of the 36th International Conference on Machine Learning*.

Zhang, L., Wang, G., Tu, W.-W., and Zhou, Z.-H. (2019b).
Dual adaptivity: A universal algorithm for minimizing the adaptive regret of convex functions.
ArXiv e-prints, arXiv:1906.10851.

Zhang, L., Yang, T., Jin, R., and Zhou, Z.-H. (2018).
Dynamic regret of strongly adaptive methods.
In *Proceedings of the 35th International Conference on Machine Learning*.

Zinkevich, M. (2003).
Online convex programming and generalized infinitesimal gradient ascent.
In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936.