LAMDA Learning And Mining from DatA

南京大學 NANJING UNIVERSITY

Online Learning in Dynamic Environment

Lijun Zhang

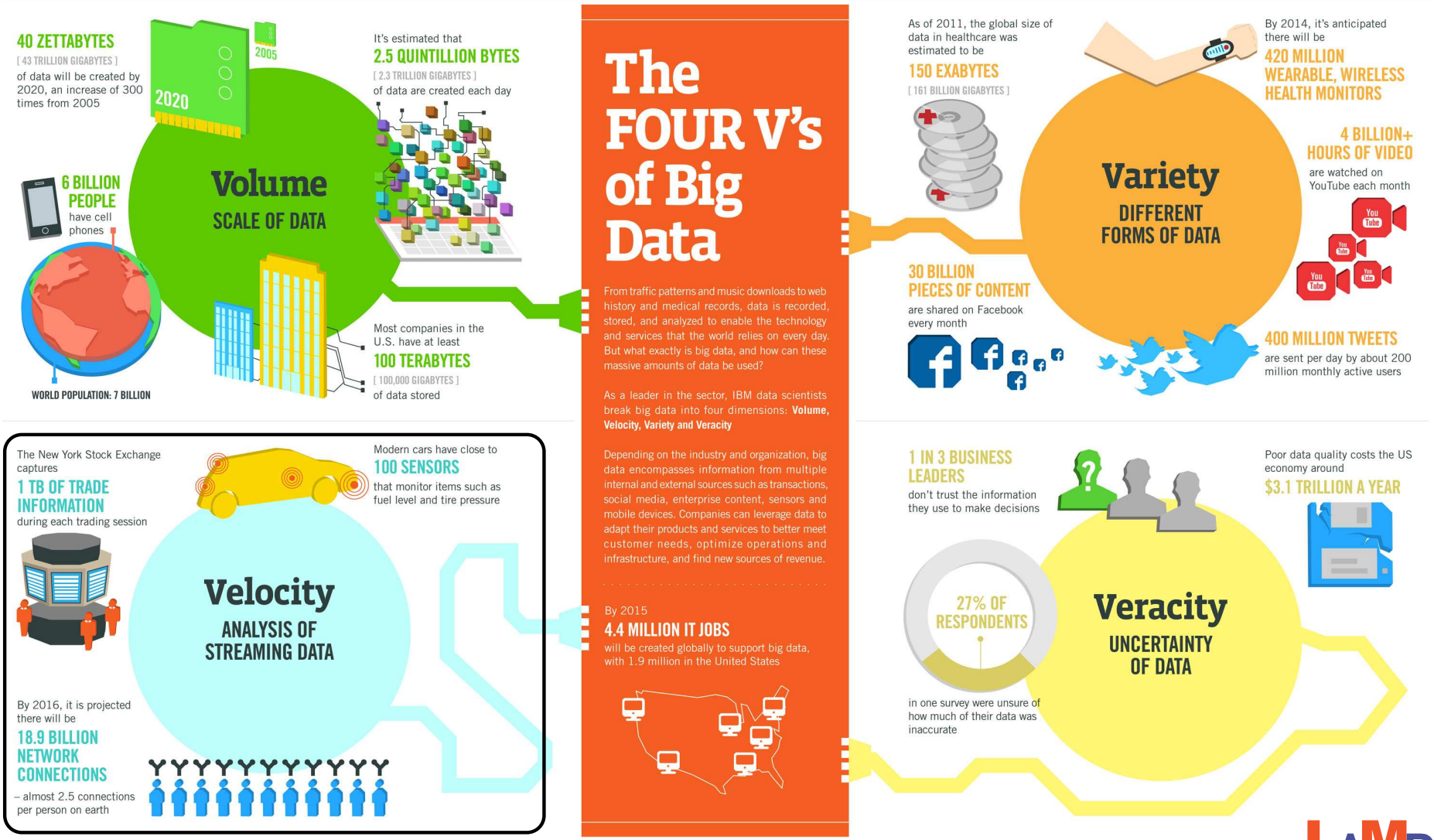
Dept. of Computer Science and Technology, Nanjing University, China

September 29, 2016

Outline

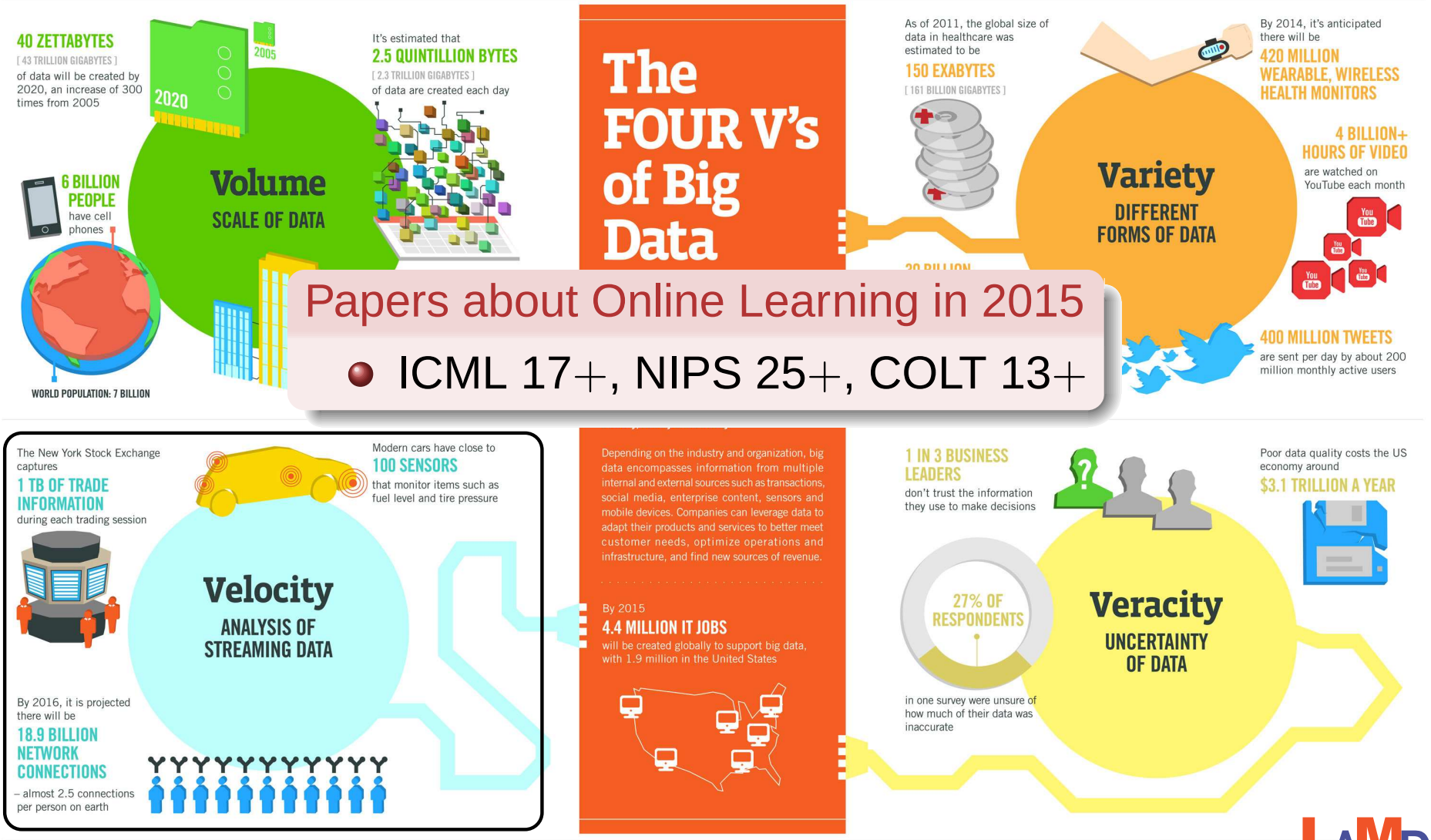
LAMDA
Learning And Mining from DatA

Big Data



http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Big Data



Papers about Online Learning in 2015

- ICML 17+, NIPS 25+, COLT 13+

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Outline

LAMDA
Learning And Mining from DatA

Online Learning

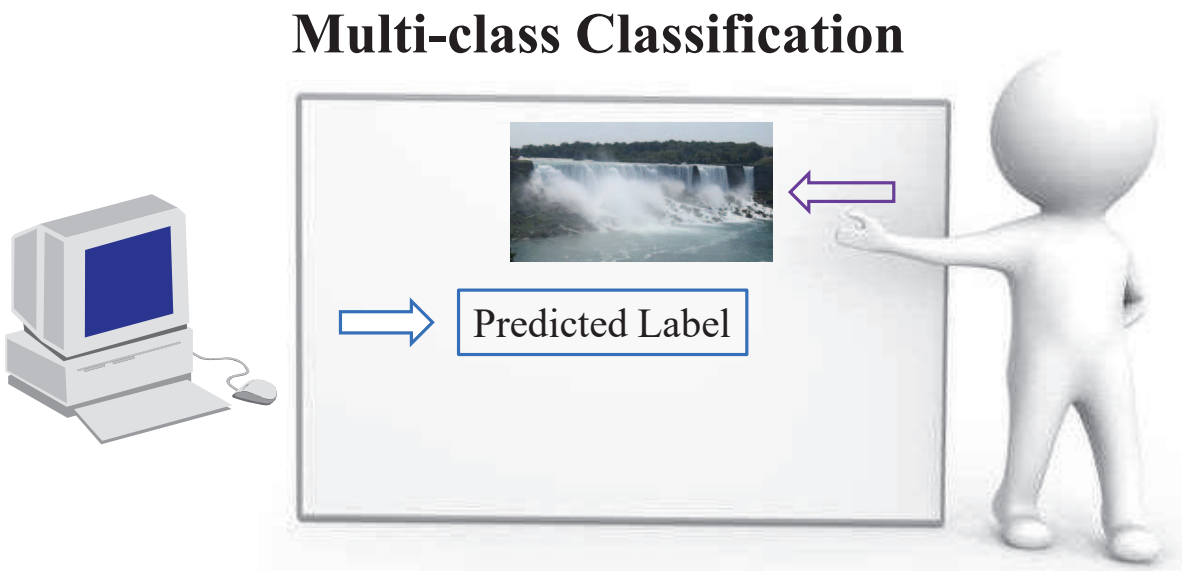
Online Learning [Shalev-Shwartz, 2011]

Online learning is the process of answering a sequence of questions given (maybe partial) knowledge of answers to previous questions and possibly additional information.

Online Learning

Online Learning [Shalev-Shwartz, 2011]

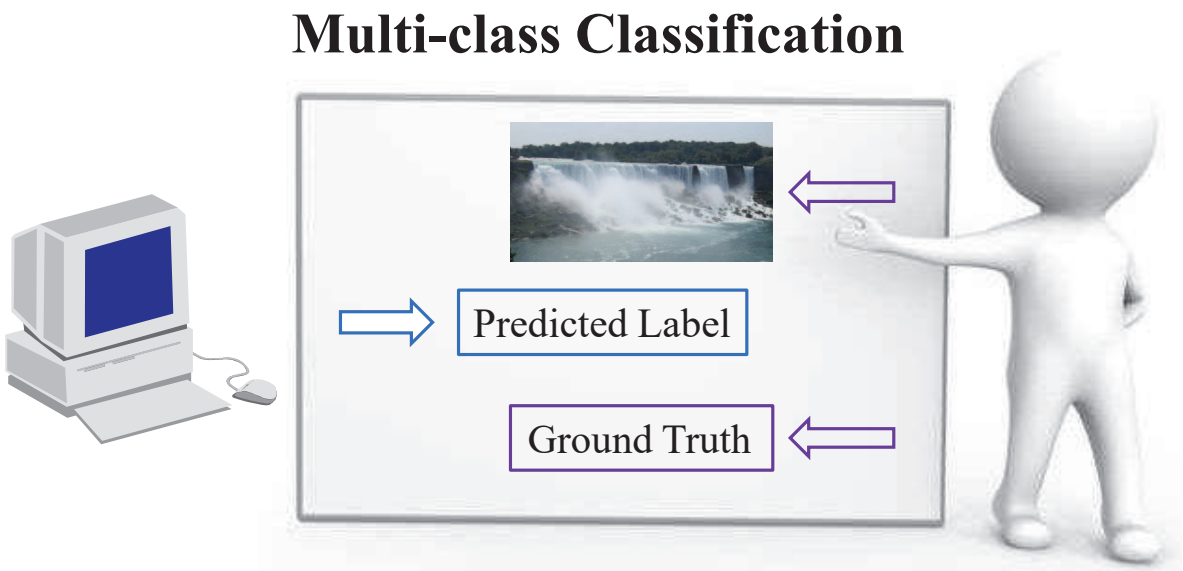
Online learning is the process of answering a sequence of questions given (maybe partial) knowledge of answers to previous questions and possibly additional information.



Online Learning

Online Learning [Shalev-Shwartz, 2011]

Online learning is the process of answering a sequence of questions given (maybe partial) knowledge of answers to previous questions and possibly additional information.

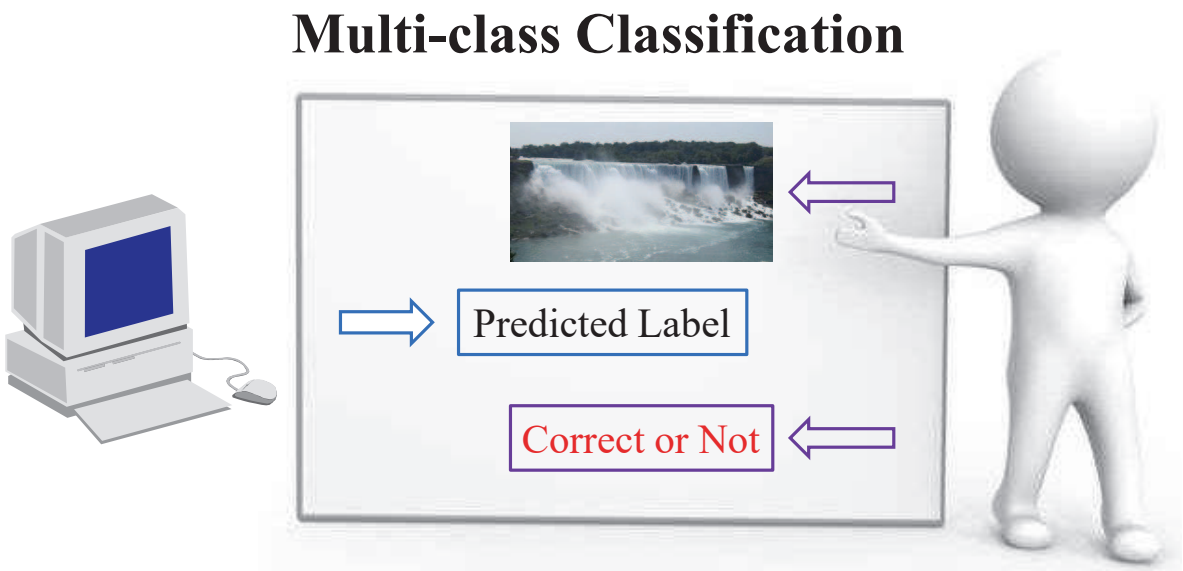


- Full-Information Online Learning

Online Learning

Online Learning [Shalev-Shwartz, 2011]

Online learning is the process of answering a sequence of questions given (maybe partial) knowledge of answers to previous questions and possibly additional information.



- Bandit Online Learning

Formal Definitions

Online Learning

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

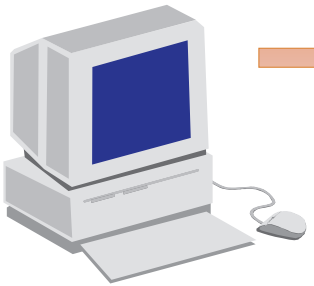
4: **end for**

Formal Definitions

Online Learning

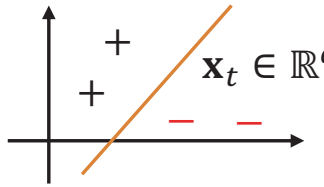
1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner picks a decision $\mathbf{x}_t \in \mathcal{X}$
Adversary chooses a function $f_t(\cdot)$

4: **end for**



Learner

A classifier



An example $(\mathbf{z}_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$

A loss $f_t(\mathbf{x}) = \max(1 - y_t \mathbf{x}^\top \mathbf{z}_t, 0)$

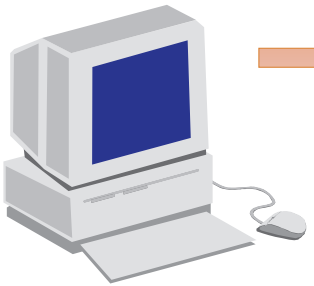
Adversary

LAMDA Learning And Mining from DatA

Formal Definitions

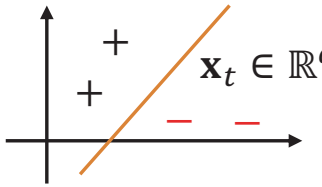
Online Learning

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



Learner

A classifier



An example $(\mathbf{z}_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$

A loss $f_t(\mathbf{x}) = \max(1 - y_t \mathbf{x}^\top \mathbf{z}_t, 0)$

Adversary

Suffer $f_t(\mathbf{x}_t)$ and update $\mathbf{x}_t$

LAMDA Learning And Mining from DatA

Formal Definitions

Online Learning

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

Cumulative Loss

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

Formal Definitions

Online Learning

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

Cumulative Loss

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

- Full-Information Online Learning
  - Learner observes the function $f_t(\cdot)$
  - Online first-order optimization
- Bandit Online Learning
  - Learner only observes the value of $f_t(\mathbf{x}_t)$
  - Online zero-order optimization

LAMDA Learning And Mining from DatA

Outline

LAMDA
Learning And Mining from DatA

Regret

Cumulative Loss

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

Regret

Cumulative Loss

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

Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

LAMDA Learning And Mining from DatA

Regret

Cumulative Loss

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

Regret

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

LAMDA Learning And Mining from DatA

Regret

Cumulative Loss

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

Regret

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

Hannan Consistent

$$\limsup_{T\to\infty} \frac{1}{T}\left(\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})\right) = 0, \text{ with probability 1}$$

LAMDA Learning And Mining from DatA

Regret

Cumulative Loss

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

Regret

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

Hannan Consistent

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) = o(T), \text{ with probability 1}$$

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (I)

- Online Gradient Descent

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

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

4: **end for**

State-of-the-Art in Full-Information Setting (I)

- Online Gradient Descent

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

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

4: **end for**

Round	$\cdots$	t
Learner	$\cdots$	$\mathbf{x}_t$
Adversary	$\cdots$	$f_t(\cdot)$
Loss	$\cdots$	$f_t(\mathbf{x}_t)$

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (I)

- Online Gradient Descent

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

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

4: **end for**

Round	$\cdots$	t	$t+1$
Learner	$\cdots$	$\mathbf{x}_t$	
Adversary	$\cdots$	$f_t(\cdot)$	$f_{t+1}(\cdot)$
Loss	$\cdots$	$f_t(\mathbf{x}_t)$	$f_{t+1}(\mathbf{x}_{t+1})$

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (I)

- Online Gradient Descent

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

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

4: **end for**

Round	$\cdots$	t	$t+1$
Learner	$\cdots$	$\mathbf{x}_t$	$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f_{t+1}(\mathbf{x})$
Adversary	$\cdots$	$f_t(\cdot)$	$f_{t+1}(\cdot)$
Loss	$\cdots$	$f_t(\mathbf{x}_t)$	$f_{t+1}(\mathbf{x}_{t+1})$

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (I)

Online Gradient Descent

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

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

4: **end for**

Round	$\cdots$	t	$t+1$
Learner	$\cdots$	$\mathbf{x}_t$	$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_{t+1}(\mathbf{x}_t) \right)$
Adversary	$\cdots$	$f_t(\cdot)$	$f_{t+1}(\cdot)$
Loss	$\cdots$	$f_t(\mathbf{x}_t)$	$f_{t+1}(\mathbf{x}_{t+1})$

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (I)

- Online Gradient Descent

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

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

4: **end for**

Round	$\cdots$	t	$t+1$
Learner	$\cdots$	$\mathbf{x}_t$	$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$
Adversary	$\cdots$	$f_t(\cdot)$	$f_{t+1}(\cdot)$
Loss	$\cdots$	$f_t(\mathbf{x}_t)$	$f_{t+1}(\mathbf{x}_{t+1})$

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (II)

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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{D^2}{2}\sqrt{T} + \left(\sqrt{T} - \frac{1}{2}\right) G^2 = O\left(\sqrt{T}\right)$$

State-of-the-Art in Full-Information Setting (II)

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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{D^2}{2}\sqrt{T} + \left(\sqrt{T} - \frac{1}{2}\right) G^2 = O\left(\sqrt{T}\right)$$

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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{G^2}{2\lambda}(1 + \log T) = O(\log T)$$

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (II)

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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{D^2}{2}\sqrt{T} + \left(\sqrt{T} - \frac{1}{2}\right) G^2 = O\left(\sqrt{T}\right)$$

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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{G^2}{2\lambda}(1 + \log T) = O(\log T)$$

- Exponentially Concave Functions [Hazan et al., 2007]
- Convex and Smooth Functions [Srebro et al., 2010]
- Sparse Online Learning [Langford et al., 2009]
- Online Classification [Freund and Schapire, 1999]

LAMDA Learning And Mining from DatA

State-of-the-Art in Full-Information Setting (II)

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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{D^2}{2}\sqrt{T} + \left(\sqrt{T} - \frac{1}{2}\right) G^2 = O\left(\sqrt{T}\right)$$

- Strongly
  - Online

Two Characteristics

- Updating only once per round
- A decreasing step size

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{G^2}{2\lambda}(1 + \log T) = O(\log T)$$

- Exponentially Concave Functions [Hazan et al., 2007]
- Convex and Smooth Functions [Srebro et al., 2010]
- Sparse Online Learning [Langford et al., 2009]
- Online Classification [Freund and Schapire, 1999]

LAMDA Learning And Mining from DatA

Outline

LAMDA
Learning And Mining from DatA

A Paradox

- The Ideal Algorithm

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_{t+1}(\mathbf{x}_t) \right)$$

- Online Gradient Descent is Imperfect

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

  - Fail if f_t and f_{t+1} differ much

A Paradox

- The Ideal Algorithm

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_{t+1}(\mathbf{x}_t) \right)$$

- Online Gradient Descent is Imperfect

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \right)$$

  - Fail if f_t and f_{t+1} differ much

- Its Theoretical Guarantees are Strong

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \begin{cases} O\left(\sqrt{T}\right), & \text{Convex Functions} \\ O\left(\log T\right), & \text{Strongly Convex Functions} \end{cases}$$

LAMDA Learning And Mining from DatA

Dynamic Regret

Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

Dynamic Regret

Regret →Static Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

- The Rationale of Regret
  - One of the decision is reasonably good during T rounds

Dynamic Regret

Regret →Static Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

- The Rationale of Regret
  - One of the decision is reasonably good during T rounds
- Dynamic Environment
  - Different decisions will be good in different periods
  - E.g. online tracking, online advertising, portfolio investment

Dynamic Regret

Regret →Static Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

- The Rationale of Regret
  - One of the decision is reasonably good during T rounds
- Dynamic Environment
  - Different decisions will be good in different periods
  - E.g. online tracking, online advertising, portfolio investment

Dynamic Regret

$$\text{Dynamic Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$$

LAMDA Learning And Mining from DatA

Dynamic Regret

Regret →Static Regret

$$\text{Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

- The Rationale of Regret
  - One of the decision is reasonably good during T rounds
- Dynamic Environment
  - Different decisions will be good in different periods
  - E.g. online tracking, online advertising, portfolio investment

Dynamic Regret

$$\text{Dynamic Regret} = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$$

LAMDA Learning And Mining from DatA

The Challenge

Sublinear Dynamic Regret is Impossible in General!

The Challenge

Sublinear Dynamic Regret is Impossible in General!

- An Example—A Random Adversary

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner picks a decision $x_t \in \mathbb{R}$
3: Adversary chooses $(x-1)^2$ and $(x+1)^2$ randomly
4: **end for**

The Challenge

Sublinear Dynamic Regret is Impossible in General!

- An Example—A Random Adversary

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner picks a decision $x_t \in \mathbb{R}$
3: Adversary chooses $(x-1)^2$ and $(x+1)^2$ randomly
4: **end for**

- The expectation of dynamic regret at t-th round

$$\frac{1}{2}\left((x_t-1)^2+(x_t+1)^2\right)-\frac{1}{2}\left(\min_x (x-1)^2+\min_x (x+1)^2\right)$$
$$=x_t^2+1\geq 1$$

The Challenge

Sublinear Dynamic Regret is Impossible in General!

- An Example—A Random Adversary

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner picks a decision $x_t \in \mathbb{R}$
3: Adversary chooses $(x-1)^2$ and $(x+1)^2$ randomly
4: **end for**

- The expectation of dynamic regret at t-th round

$$\frac{1}{2}\left((x_t-1)^2+(x_t+1)^2\right)-\frac{1}{2}\left(\min_x (x-1)^2+\min_x (x+1)^2\right)$$
$$=x_t^2+1\geq 1$$

- $$\mathrm{E}\left[\sum_{t=1}^{T} f_t(x_t)-\sum_{t=1}^{T}\min_x f_t(x)\right]\geq T$$

Make the Impossible Possible

Dynamic Regret

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^*)$$

where $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$.

Make the Impossible Possible

Dynamic Regret

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^*)$$

where $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$.

- Assumptions
  - Functional Variation [Besbes et al., 2015]

$$\mathcal{F}_T := \sum_{t=2}^{T} \|f_t(\cdot) - f_{t-1}(\cdot)\|_\infty = \sum_{t=2}^{T} \max_{\mathbf{x}\in\mathcal{X}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})|$$

  - Path-length [Mokhtari et al., 2016]

$$\mathcal{P}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|$$

LAMDA Learning And Mining from DatA

Existing Results (I)

- Functional Variation [Besbes et al., 2015]
  - For any $f_1, \ldots, f_T$ such that

$$\mathcal{F}_T := \sum_{t=2}^{T} \max_{\mathbf{x} \in \mathcal{X}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})| \leq V_T$$

    the restarted online gradient descent (with input V_T) satisfies

$$\text{Dynamic Regret} \leq \begin{cases} O\left(T^{2/3} {V_T}^{1/3}\right), \text{ Convex Functions} \\ O\left(\log T \sqrt{T V_T}\right), \text{ Strongly Convex Functions} \end{cases}$$

  - Dynamic regret is sublinear if V_T is sublinear

Existing Results (I)

- Functional Variation [Besbes et al., 2015]
  - For any $f_1, \ldots, f_T$ such that

$$\mathcal{F}_T := \sum_{t=2}^{T} \max_{\mathbf{x} \in \mathcal{X}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})| \leq V_T$$

    the restarted online gradient descent (with input V_T) satisfies

$$\text{Dynamic Regret} \leq \begin{cases} O\left(T^{2/3} {V_T}^{1/3}\right), \text{ Convex Functions} \\ O\left(\log T \sqrt{T V_T}\right), \text{ Strongly Convex Functions} \end{cases}$$

  - Dynamic regret is sublinear if V_T is sublinear
  - It is not adaptive

Existing Results (II)

- Path-length

[Zinkevich, 2003, Mokhtari et al., 2016, Yang et al., 2016]

- Online gradient descent satisfies

$$\text{Dynamic Regret} \leq \begin{cases} O\left(\sqrt{T\mathcal{P}_T^*}\right), \text{ Convex Functions} \\ O\left(\mathcal{P}_T^*\right), \text{ Strongly Convex and Smooth} \\ O\left(\mathcal{P}_T^*\right), \text{ Convex and Smooth Functions} \end{cases}$$

- For any $f_1, \ldots, f_T$ such that

$$\mathcal{P}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\| \leq B_T$$

online gradient descent (with input B_T) satisfies

$$\text{Dynamic Regret} \leq O\left(\sqrt{TB_T}\right), \text{ Convex Functions}$$

LAMDA Learning And Mining from DatA

Our Contribution [Zhang et al., 2016]

- Squared Path-length

$$\mathcal{S}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|^2$$

Our Contribution [Zhang et al., 2016]

- Squared Path-length

$$\mathcal{S}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|^2$$

Example 1 (Path-length vs Squared Path-length)

Suppose $\|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\| = T^{-\alpha}$ *for all* $t \geq 1$ *and* $\alpha > 0$, *we have*

$$\mathcal{S}_{T+1}^* = T^{1-2\alpha} \ll \mathcal{P}_{T+1}^* = T^{1-\alpha}.$$

In particular, when $\alpha = 1/2$, *we have*

$$\mathcal{S}_{T+1}^* = 1 \ll \mathcal{P}_{T+1}^* = \sqrt{T}.$$

Our Contribution [Zhang et al., 2016]

- Squared Path-length

$$\mathcal{S}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|^2$$

- Online Multiple Gradient Descent
  - Strongly convex and smooth functions
  - Semi-strongly convex and smooth functions
  -

$$\text{Dynamic Regret} \leq O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right)$$

LAMDA Learning And Mining from DatA

Our Contribution [Zhang et al., 2016]

- Squared Path-length

$$\mathcal{S}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|^2$$

- Online Multiple Gradient Descent
  - Strongly convex and smooth functions
  - Semi-strongly convex and smooth functions
  - $$\text{Dynamic Regret} \leq O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right)$$

- Online Multiple Newton Update
  - Self-concordant functions
  - $$\text{Dynamic Regret} \leq O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right)$$

LAMDA Learning And Mining from DatA

Outline

LAMDA
Learning And Mining from DatA

Strongly Convex and Smooth Functions

Definition 1

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is λ-strongly convex, if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Definition 2

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is L-smooth, if

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

LAMDA Learning And Mining from DatA

Strongly Convex and Smooth Functions

Definition 1

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is λ-strongly convex, if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Definition 2

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is L-smooth, if

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Example 2 (Strongly convex and smooth functions)

- *A quadratic form* $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} - 2\mathbf{b}^\top \mathbf{x} + c$ *where* $\alpha I \preceq \mathcal{A} \preceq \beta I$, $\alpha > 0$ *and* $\beta < \infty$;
- *The regularized logistic loss* $f(\mathbf{x}) = \log(1 + \exp(\mathbf{b}^\top \mathbf{x})) + \frac{\lambda}{2}\|\mathbf{x}\|^2$, *where* $\lambda > 0$.

LAMDA Learning And Mining from DatA

Online Multiple Gradient Descent

- The Algorithm

Let $\mathbf{x}_1$ be any point in $\mathcal{X}$
for $t = 1, \ldots, T$ **do**
    $\mathbf{z}_t^1 = \mathbf{x}_t$
    for $j = 1, \ldots, K$ **do**

$$\mathbf{z}_t^{j+1} = \Pi_{\mathcal{X}} \left(\mathbf{z}_t^j - \eta \nabla f_t(\mathbf{z}_t^j) \right)$$

    end for
    $\mathbf{x}_{t+1} = \mathbf{z}_t^{K+1}$
end for

Online Multiple Gradient Descent

- The Algorithm

1: Let $\mathbf{x}_1$ be any point in $\mathcal{X}$
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad \mathbf{z}_t^1 = \mathbf{x}_t$
4: $\quad$ **for** $j = 1, \ldots, K$ **do**
5:

$$\mathbf{z}_t^{j+1} = \Pi_{\mathcal{X}} \left(\mathbf{z}_t^j - \eta \nabla f_t(\mathbf{z}_t^j) \right)$$

6: $\quad$ **end for**
7: $\quad \mathbf{x}_{t+1} = \mathbf{z}_t^{K+1}$
8: **end for**

- Multiple updatings in each round
- A constant step size

LAMDA Learning And Mining from DatA

Theoretical Guarantees

Theorem 1

By setting $\eta \leq 1/L$ *and* $K = \lceil \frac{1/\eta+\lambda}{2\lambda} \ln 4 \rceil$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)$$

$$\leq \min \begin{cases} 2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \mathbf{x}_1^*\|, \\ \frac{1}{2\alpha} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^*)\|^2 + 2(L+\alpha)\mathcal{S}_T^* + (L+\alpha)\|\mathbf{x}_1 - \mathbf{x}_1^*\|^2 \end{cases}$$

Theoretical Guarantees

Theorem 1

By setting $\eta \leq 1/L$ *and* $K = \lceil \frac{1/\eta+\lambda}{2\lambda} \ln 4 \rceil$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)$$

$$\leq \min \begin{cases} 2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \mathbf{x}_1^*\|, \\ \frac{1}{2\alpha} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^*)\|^2 + 2(L+\alpha)\mathcal{S}_T^* + (L+\alpha)\|\mathbf{x}_1 - \mathbf{x}_1^*\|^2 \end{cases}$$

Corollary 1

Suppose $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^*)\|^2 \leq O(\mathcal{S}_T^*)$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right).$$

LAMDA Learning And Mining from DatA

Theoretical Guarantees

Theorem 1

By setting $\eta \leq 1/L$ *and* $K = \lceil \frac{1/\eta+\lambda}{2\lambda} \ln 4 \rceil$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)$$

$$\leq \min \begin{cases} 2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \mathbf{x}_1^*\|, \\ \frac{1}{2\alpha} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t^*)\|^2 + 2(L+\alpha)\mathcal{S}_T^* + (L+\alpha)\|\mathbf{x}_1 - \mathbf{x}_1^*\|^2 \end{cases}$$

Corollary 1

In particular, if $\mathbf{x}_t^*$ *belongs to the relative interior of* $\mathcal{X}$

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)$$

$$\leq \min \left(2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \mathbf{x}_1^*\|, 2L\mathcal{S}_T^* + L\|\mathbf{x}_1 - \mathbf{x}_1^*\|^2 \right)$$

LAMDA Learning And Mining from DatA

Semi-strongly Convex and Smooth Functions

Definition 3 ([Gong and Ye, 2014])

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is semi-strongly convex over $\mathcal{X}$, if

$$f(\mathbf{x}) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \geq \frac{\beta}{2} \left\| \mathbf{x} - \Pi_{\mathcal{X}^*}(\mathbf{x}) \right\|^2$$

where $\mathcal{X}^* = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \leq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})\}$.

Semi-strongly Convex and Smooth Functions

Definition 3 ([Gong and Ye, 2014])

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is semi-strongly convex over $\mathcal{X}$, if

$$f(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \geq \frac{\beta}{2} \|\mathbf{x} - \Pi_{\mathcal{X}^*}(\mathbf{x})\|^2$$

where $\mathcal{X}^* = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \leq \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})\}$.

Example 3 (Semi-strongly Convex and Smooth Functions)

$$\min_{\mathbf{x}\in\mathcal{X}\subseteq\mathbb{R}^d} \quad f(\mathbf{x}) = g(E\mathbf{x}) + \mathbf{b}^\top \mathbf{x}$$

where $g(\cdot)$ is strongly convex and smooth, and $\mathcal{X}$ is either $\mathbb{R}^d$ or a polyhedral set. [Luo and Tseng, 1993]

- The semi-strong convexity and error bound property are equivalent up to constant factors [Necoara et al., 2015]

Theoretical Guarantees

Theorem 2

By setting $\eta \leq 1/L$ *and* $K = \lceil \frac{1/\eta+\beta}{\beta} \ln 4 \rceil$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} \min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$$

$$\leq \min \begin{cases} 2G\mathcal{P}_T^* + 2G\|\mathbf{x}_1 - \bar{\mathbf{x}}_1\|, \\ \frac{1}{2\alpha} G_T^* + 2(L+\alpha)\mathcal{S}_T^* + (L+\alpha)\|\mathbf{x}_1 - \bar{\mathbf{x}}_1\|^2 \end{cases}$$

where $G_T^* = \max_{\{\mathbf{x}_t^* \in \mathcal{X}_t^*\}_{t=1}^T} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t^*)\|^2$, *and* $\bar{\mathbf{x}}_1 = \Pi_{\mathcal{X}_1^*}(\mathbf{x}_1)$.

Corollary 2

Suppose $G_T^* \leq O(\mathcal{S}_T^*)$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right).$$

LAMDA Learning And Mining from DatA

Outline

LAMDA
Learning And Mining from DatA

Self-concordant Functions

Definition 4 ([Nemirovski, 2004])

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is called self-concordant on $\mathcal{X}$, if

- $f(\mathbf{x}_i) \to \infty$ along every sequence $\{\mathbf{x}_i \in \mathcal{X}\}$ converging, as $i \to \infty$, to a boundary point of $\mathcal{X}$;
- f satisfies the differential inequality

$$|D^3 f(\mathbf{x})[\mathbf{h}, \mathbf{h}, \mathbf{h}]| \leq 2 \left(\mathbf{h}^\top \nabla^2 f(\mathbf{x}) \mathbf{h}\right)^{3/2}$$

for all $\mathbf{x} \in \mathcal{X}$ and all $\mathbf{h} \in \mathbb{R}^d$.

Example 4 (Self-concordant Functions)

- *The function* $f(x) = -\log x$
- *A convex quadratic form* $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} - 2\mathbf{b}^\top \mathbf{x} + c$ *where* $A \in \mathbb{R}^{d \times d}$, $\mathbf{b} \in \mathbb{R}^d$, *and* $c \in \mathbb{R}$
- *If* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *is self-concordant, and* $A \in \mathbb{R}^{d \times k}$, $\mathbf{b} \in \mathbb{R}^d$, *then* $f(A\mathbf{x} + \mathbf{b})$ *is self-concordant.*

LAMDA Learning And Mining from DatA

Online Multiple Newton Update

- The Algorithm

Let $\mathbf{x}_1$ be any point in $\mathcal{X}_1$
for $t = 1, \ldots, T$ **do**
  $\mathbf{z}_t^1 = \mathbf{x}_t$
  for $j = 1, \ldots, K$ **do**

$$\mathbf{z}_t^{j+1} = \mathbf{z}_t^j - \frac{1}{1 + \lambda_t(\mathbf{z}_t^j)} \left[\nabla^2 f_t(\mathbf{z}_t^j)\right]^{-1} \nabla f_t(\mathbf{z}_t^j)$$

$$\lambda_t(\mathbf{z}_t^j) = \sqrt{\nabla f_t(\mathbf{z}_t^j)^\top \left[\nabla^2 f_t(\mathbf{z}_t^j)\right]^{-1} \nabla f_t(\mathbf{z}_t^j)}$$

  end for
  $\mathbf{x}_{t+1} = \mathbf{z}_t^{K+1}$
end for

LAMDA Learning And Mining from DatA

Theoretical Guarantees

Theorem 3

Assume

$$\|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|_t^2 \leq \frac{1}{144}, \ \forall t \geq 2.$$

When $t = 1$, *we choose* $K = O(1)(f_1(\mathbf{x}_1) - f_1(\mathbf{x}_1^*) + \log\log\mu)$ *such that*

$$\|\mathbf{x}_2 - \mathbf{x}_1^*\|_1^2 \leq \frac{1}{144\mu}.$$

For $t \geq 2$, *we set* $K = \lceil \log_4(16\mu) \rceil$ *then*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq \min\left(\frac{1}{6}\mathcal{P}_T^*, 4\mathcal{S}_T^* + \frac{1}{36}\right) + f_1(\mathbf{x}_1) - f_1(\mathbf{x}_1^*).$$

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right).$$

LAMDA Learning And Mining from DatA

Conclusion and Future Work

Conclusion

- Squared path-length: $\mathcal{S}_T^* := \sum_{t=2}^{T} \|\mathbf{x}_t^* - \mathbf{x}_{t-1}^*\|^2$
- Online multiple gradient descent (strongly/Semi-strongly convex and smooth functions)

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \leq O\left(\min(\mathcal{P}_T^*, \mathcal{S}_T^*)\right)$$

- Online multiple Newton update (self-concordant functions)

Future Work

- A deep understanding of functional variation, path-length, squared path-length
- Beyond dynamic regret
- Static regret of semi-strongly convex functions

LAMDA
Learning And Mining from DatA

Reference I

Thanks!

Besbes, O., Gur, Y., and Zeevi, A. (2015).
Non-stationary stochastic optimization.
Operations Research, 63(5):1227–1244.

Freund, Y. and Schapire, R. E. (1999).
Large margin classification using the perceptron algorithm.
Machine Learning, 37(3):277–296.

Gong, P. and Ye, J. (2014).
Linear convergence of variance-reduced stochastic gradient without strong convexity.
ArXiv e-prints, arXiv:1406.1102.

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

Langford, J., Li, L., and Zhang, T. (2009).
Sparse online learning via truncated gradient.
Journal of Machine Learning Research, 10:777–801.

Luo, Z.-Q. and Tseng, P. (1993).
Error bounds and convergence analysis of feasible descent methods: a general approach.
Annals of Operations Research, 46-47(1):157–178.

Mokhtari, A., Shahrampour, S., Jadbabaie, A., and Ribeiro, A. (2016).
Online optimization in dynamic environments: Improved regret rates for strongly convex problems.
ArXiv e-prints, arXiv:1603.04954.

LAMDA
Learning And Mining from DatA

Reference II

Necoara, I., Nesterov, Y., and Glineur, F. (2015).
Linear convergence of first order methods for non-strongly convex optimization.
ArXiv e-prints, arXiv:1504.06298.

Nemirovski, A. (2004).
Interior point polynomial time methods in convex programming.
Lecture notes, Technion – Israel Institute of Technology.

Shalev-Shwartz, S. (2011).
Online learning and online convex optimization.
Foundations and Trends in Machine Learning, 4(2):107–194.

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.

Yang, T., Zhang, L., Jin, R., and Yi, J. (2016).
Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient.
In *Proceedings of the 33rd International Conference on Machine Learning*.

Zhang, L., Yang, T., Yi, J., Jin, R., and Zhou, Z.-H. (2016).
Improved dynamic regret for non-degeneracy functions.
ArXiv e-prints, arXiv:1608.03933.

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

LAMDA
Learning And Mining from DatA