Online Stochastic Linear Optimization under One-bit Feedback

Lijun Zhang

LAMDA group, Nanjing University, China

The 6th Vision And Learning SEminar (VALSE 2016)

Outline

LAMDA
Learning And Mining from DatA

What Happens in an Internet Minute?

http://www.dailyinfographic.com/what-happens-in-an-internet-minute-infographic

Outline

LAMDA Learning And Mining from DatA

Online Algorithm vs Online Learning

Online Algorithm [Karp, 1992]

An online algorithm is one that receives a sequence of requests and performs an immediate action in response to each request.

- Computer Vision, Machine Learning, Data Mining
- Theoretical Computer Science, Computer Networks

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.

- Machine Learning, Game Theory, Information Theory

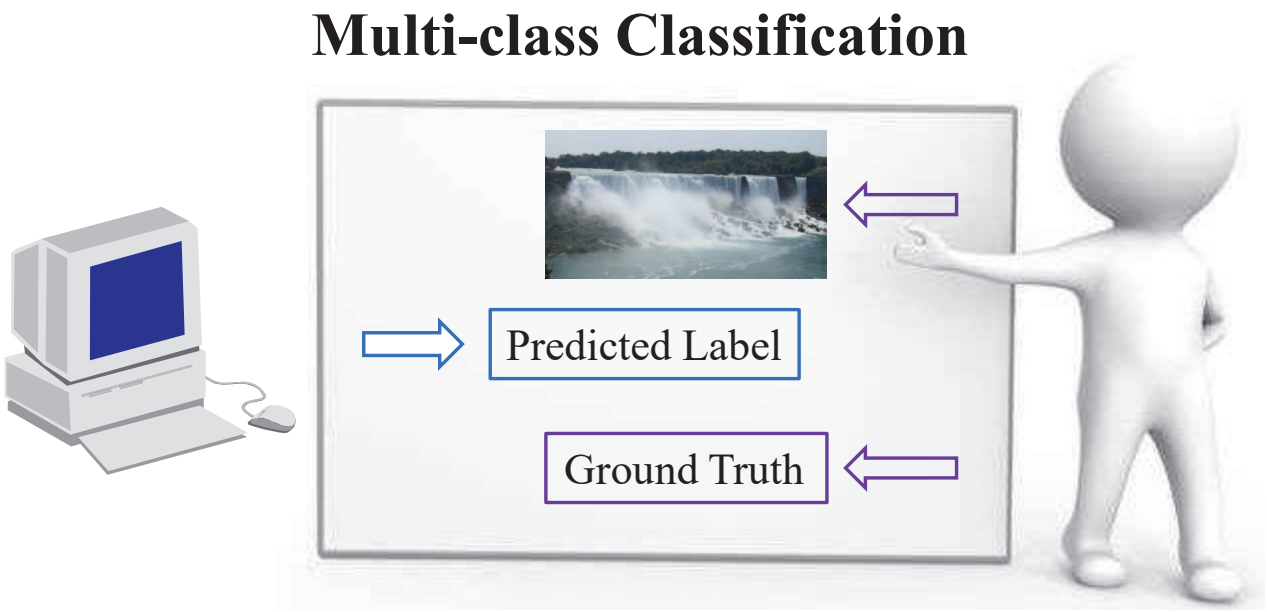
LAMDA
Learning And Mining from DatA

Full-Information vs Bandit

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



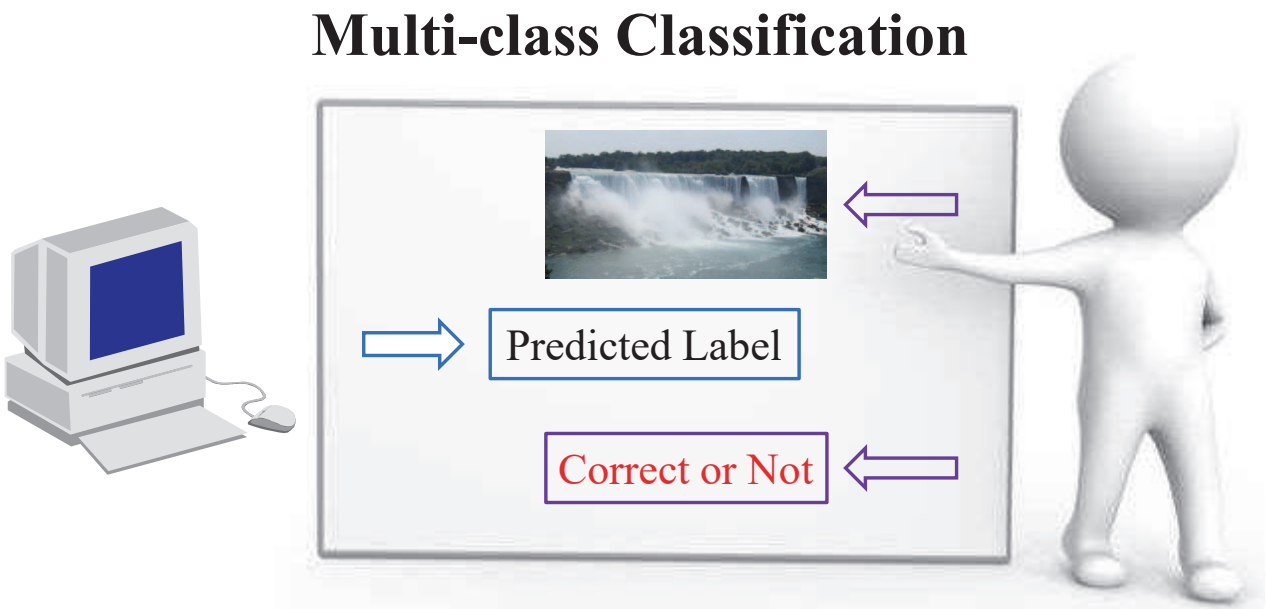
- Bandit Online Learning

Full-Information vs Bandit

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



- 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 $\mathbf{x}_t \in \mathcal{D}$
Adversary chooses a function $f_t(\cdot)$

4: **end for**

Formal Definitions

Online Learning

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

Formal Definitions

Online Learning

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

Cumulative Loss

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

Formal Definitions

Online Learning

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

Regret

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

LAMDA Learning And Mining from DatA

Formal Definitions

Online Learning

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

Regret

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

- Full-Information Online Learning
  - Learner observes the function $f_t(\cdot)$
    [Zhang et al., 2012, Zhang et al., 2013]
- Bandit Online Learning
  - Learner only observes the value of $f_t(\mathbf{x}_t)$
    [Zhang et al., 2015, Zhang et al., 2016]

LAMDA Learning And Mining from DatA

Outline

LAMDA Learning And Mining from DatA

Bandit Online Learning

- Learner observes the value of $f_t(\mathbf{x}_t)$ sequentially
  - Offline Counterpart: zero-order optimization

Bandit Online Learning

- Learner observes the value of $f_t(\mathbf{x}_t)$ sequentially
  - Offline Counterpart: zero-order optimization

- Learning Scenarios
  - Multi-Armed Bandits (MAB) [Robbins, 1952]
  - Multi-class Classification with Bandit Feedback
  - Online Convex Optimization with Bandit Feedback
    - Linear Bandits

Bandit Online Learning

- Learner observes the value of $f_t(\mathbf{x}_t)$ sequentially
  - Offline Counterpart: zero-order optimization
- Learning Scenarios
  - Multi-Armed Bandits (MAB) [Robbins, 1952]
  - Multi-class Classification with Bandit Feedback
  - Online Convex Optimization with Bandit Feedback
    - Linear Bandits
- Generation Process of f_t's
  - Stochastic: $f_1, \ldots, f_t$ are i.i.d.
  - Adversarial
    - Oblivious: f_t is independent of $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$ (like exam)
    - Nonoblivious: f_t depends on $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$ (like interview)

LAMDA Learning And Mining from DatA

Bandit Online Learning

- Learner observes the value of $f_t(\mathbf{x}_t)$ sequentially
  - Offline Counterpart: zero-order optimization
- Learning Scenarios
  - Multi-Armed Bandits (MAB) [Robbins, 1952]
  - Multi-class Classification with Bandit Feedback
  - Online Convex Optimization with Bandit Feedback
    - Linear Bandits
- Generation Process of f_t's
  - Stochastic: $f_1, \ldots, f_t$ are i.i.d.
  - Adversarial
    - Oblivious: f_t is independent of $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$ (like exam)
    - Nonoblivious: f_t depends on $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$ (like interview)

LAMDA Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

LAMDA
Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

    Arm 1
    Arm 2
    Arm 3

LAMDA
Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$
Arm 2	$X_{2,1}$
Arm 3	$X_{3,1}$

LAMDA
Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$
Arm 2	10
Arm 3	$X_{3,1}$

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$
Arm 2	10	$X_{2,2}$
Arm 3	$X_{3,1}$	$X_{3,2}$

LAMDA
Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$
Arm 2	10	$X_{2,2}$
Arm 3	$X_{3,1}$	0

LAMDA
Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$
Arm 2	10	$X_{2,2}$	$X_{2,3}$
Arm 3	$X_{3,1}$	0	$X_{3,3}$

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$	6
Arm 2	10	$X_{2,2}$	$X_{2,3}$
Arm 3	$X_{3,1}$	0	$X_{3,3}$

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$	6	$X_{1,4}$
Arm 2	10	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$
Arm 3	$X_{3,1}$	0	$X_{3,3}$	$X_{3,4}$

LAMDA
Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$	6	$X_{1,4}$
Arm 2	10	$X_{2,2}$	$X_{2,3}$	0
Arm 3	$X_{3,1}$	0	$X_{3,3}$	$X_{3,4}$

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$	6	$X_{1,4}$	$X_{1,5}$
Arm 2	10	$X_{2,2}$	$X_{2,3}$	0	$X_{2,5}$
Arm 3	$X_{3,1}$	0	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$

LAMDA
Learning And Mining from DatA

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$	6	$X_{1,4}$	$X_{1,5}$
Arm 2	10	$X_{2,2}$	$X_{2,3}$	0	$X_{2,5}$
Arm 3	$X_{3,1}$	0	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$

  - **Exploration** vs **Exploitation**

Stochastic Multi-Armed Bandits (MAB)

- Multi-Armed Bandits (MAB)
  - A gambler is facing K arms, and each time he pulls 1 arm and receives a reward

Arm 1	$X_{1,1}$	$X_{1,2}$	6	$X_{1,4}$	$X_{1,5}$
Arm 2	10	$X_{2,2}$	$X_{2,3}$	0	$X_{2,5}$
Arm 3	$X_{3,1}$	0	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$

  - **Exploration** vs **Exploitation**

- Stochastic Setting
  - Rewards of the i-th arm are i.i.d. with unknown mean μ_i

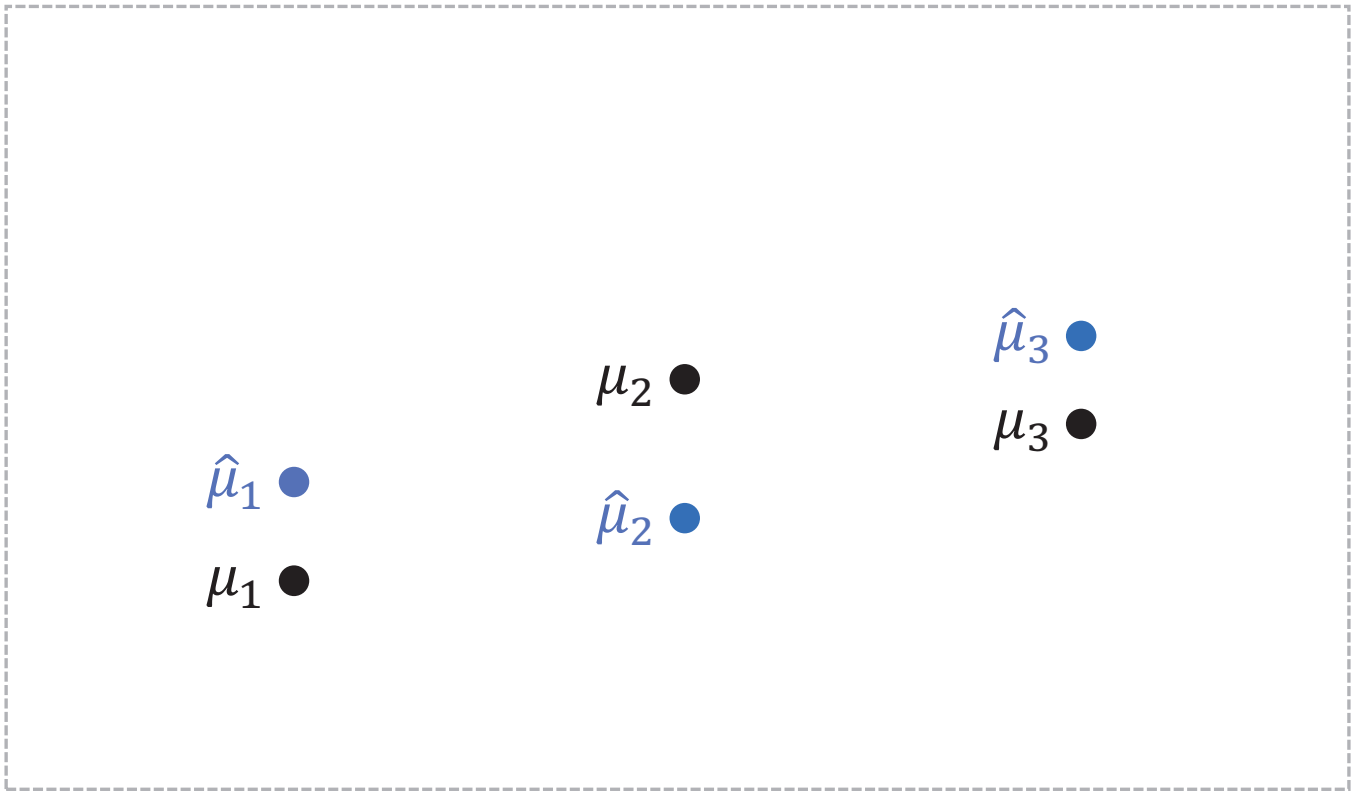
$$\text{Regret} = T \max_{i \in [K]} \mu_i - \sum_{t=1}^{T} \mu_{i_t}$$

  - Upper Confidence Bound (UCB) [Auer et al., 2002]

LAMDA Learning And Mining from DatA

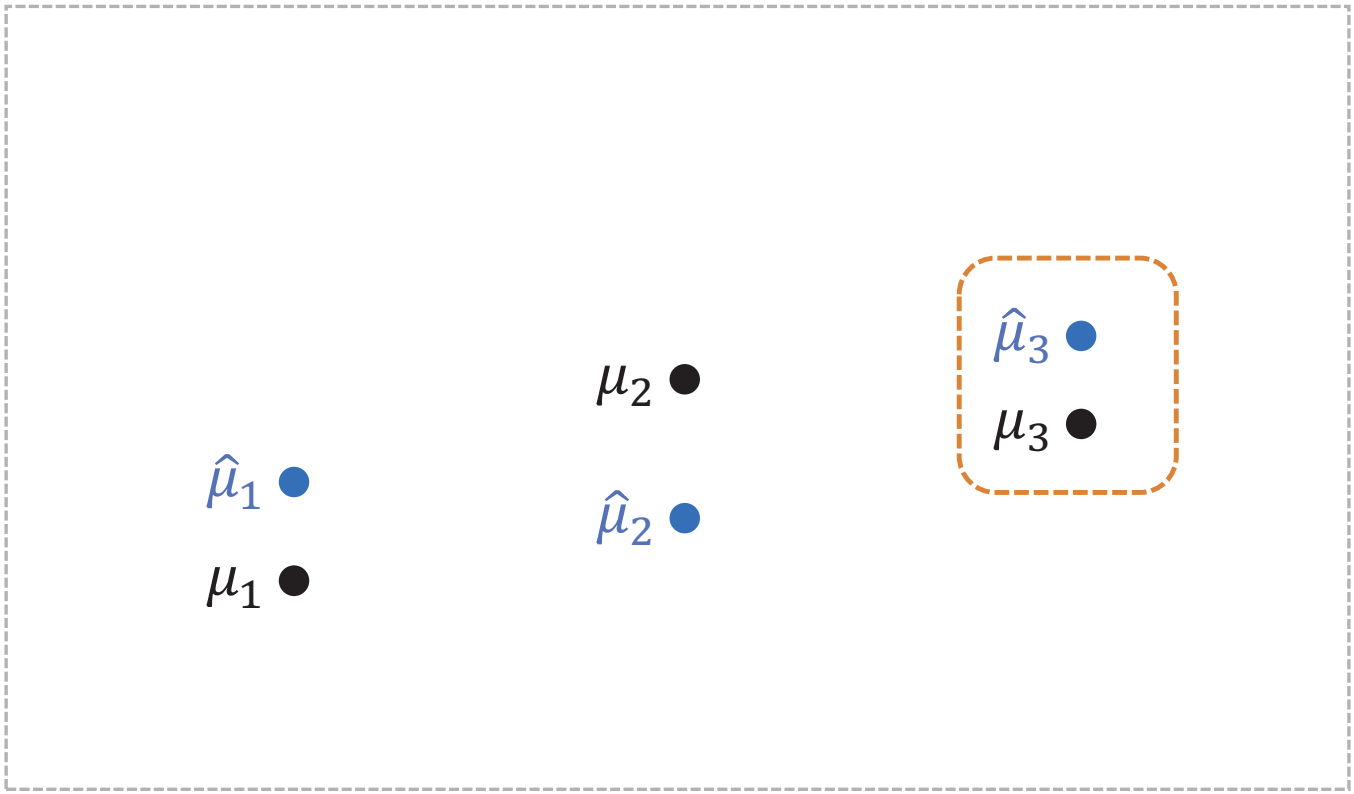
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



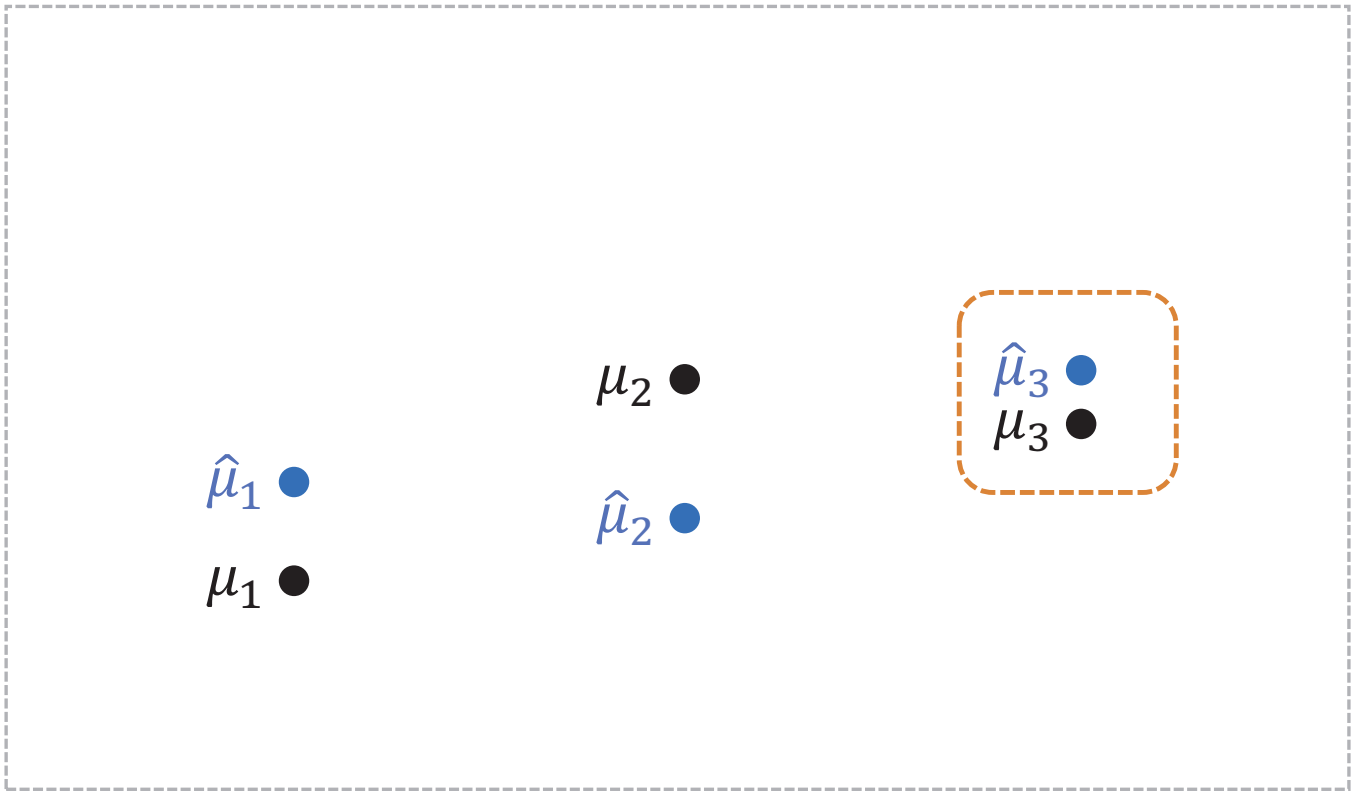
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



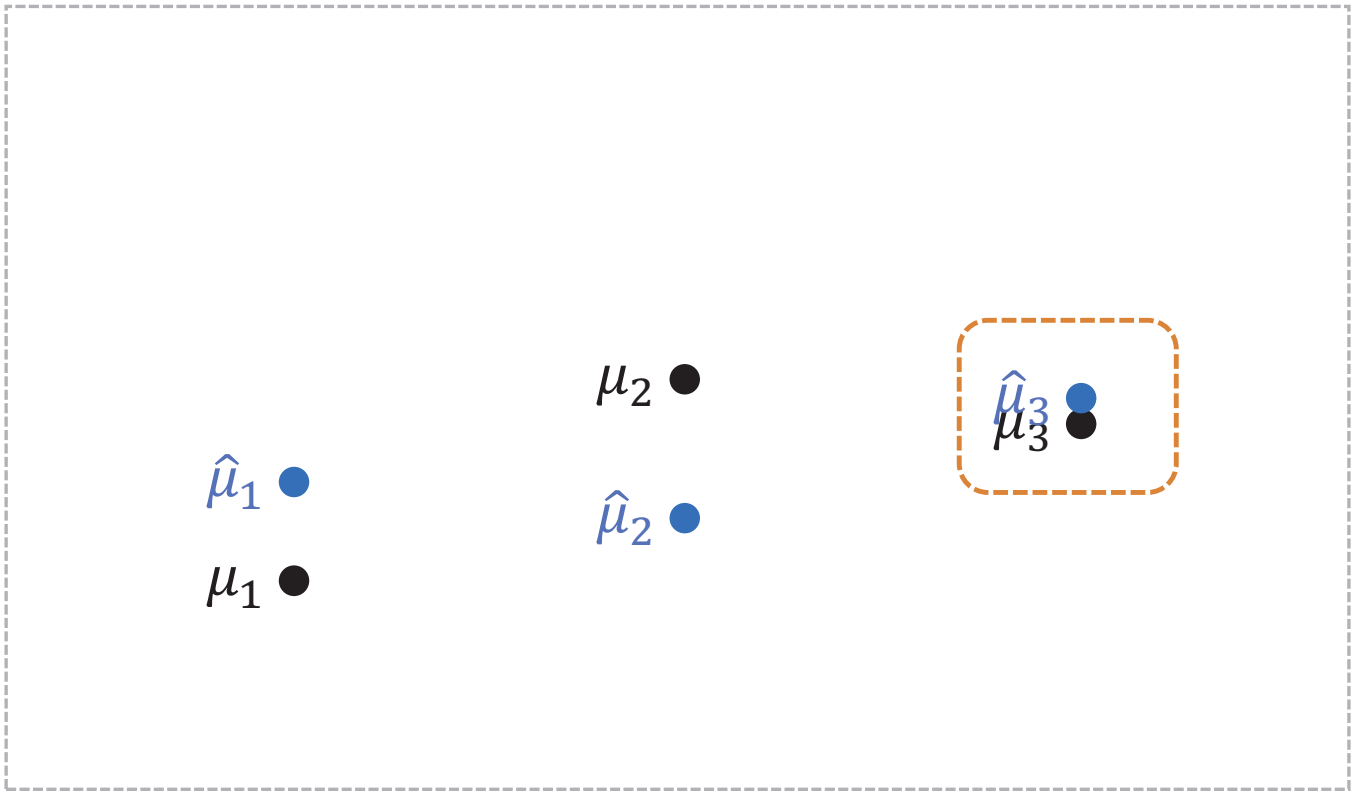
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



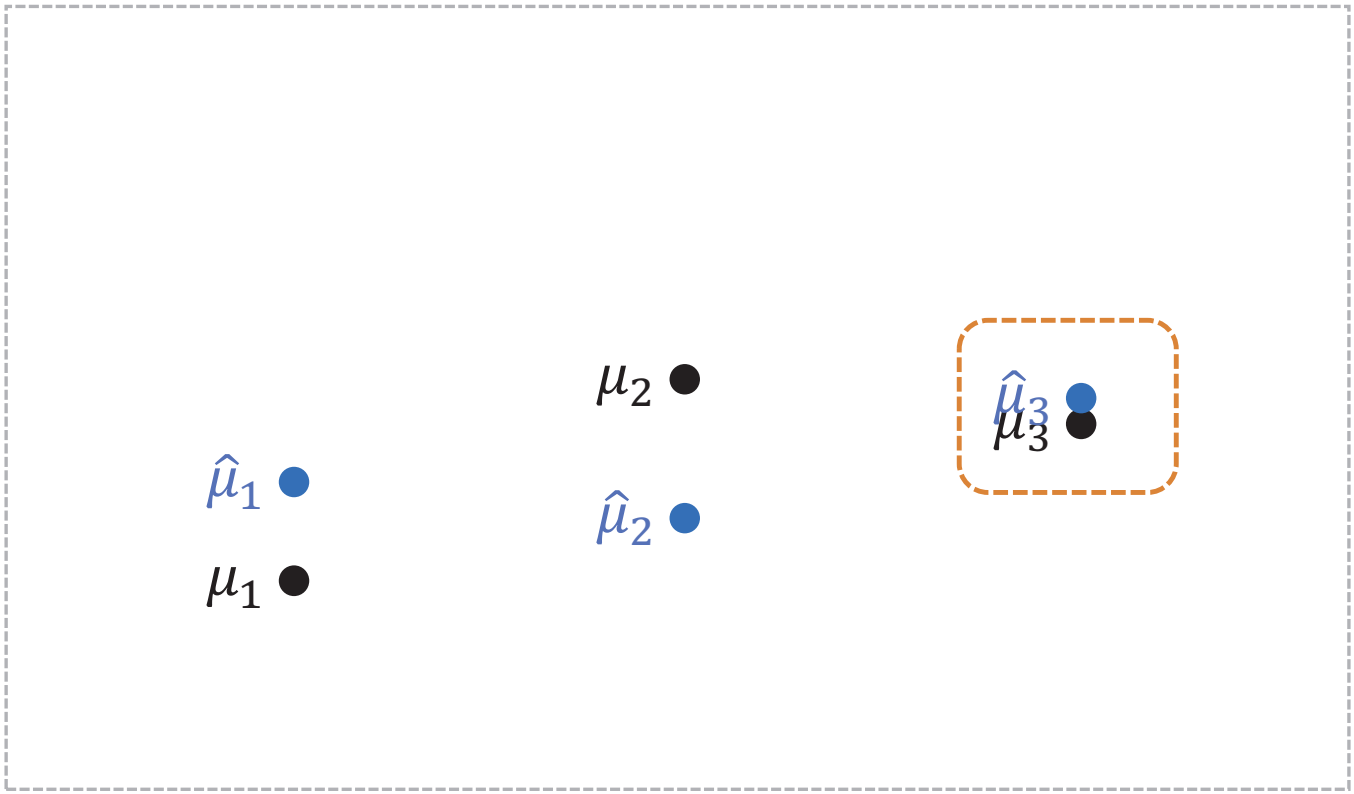
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean

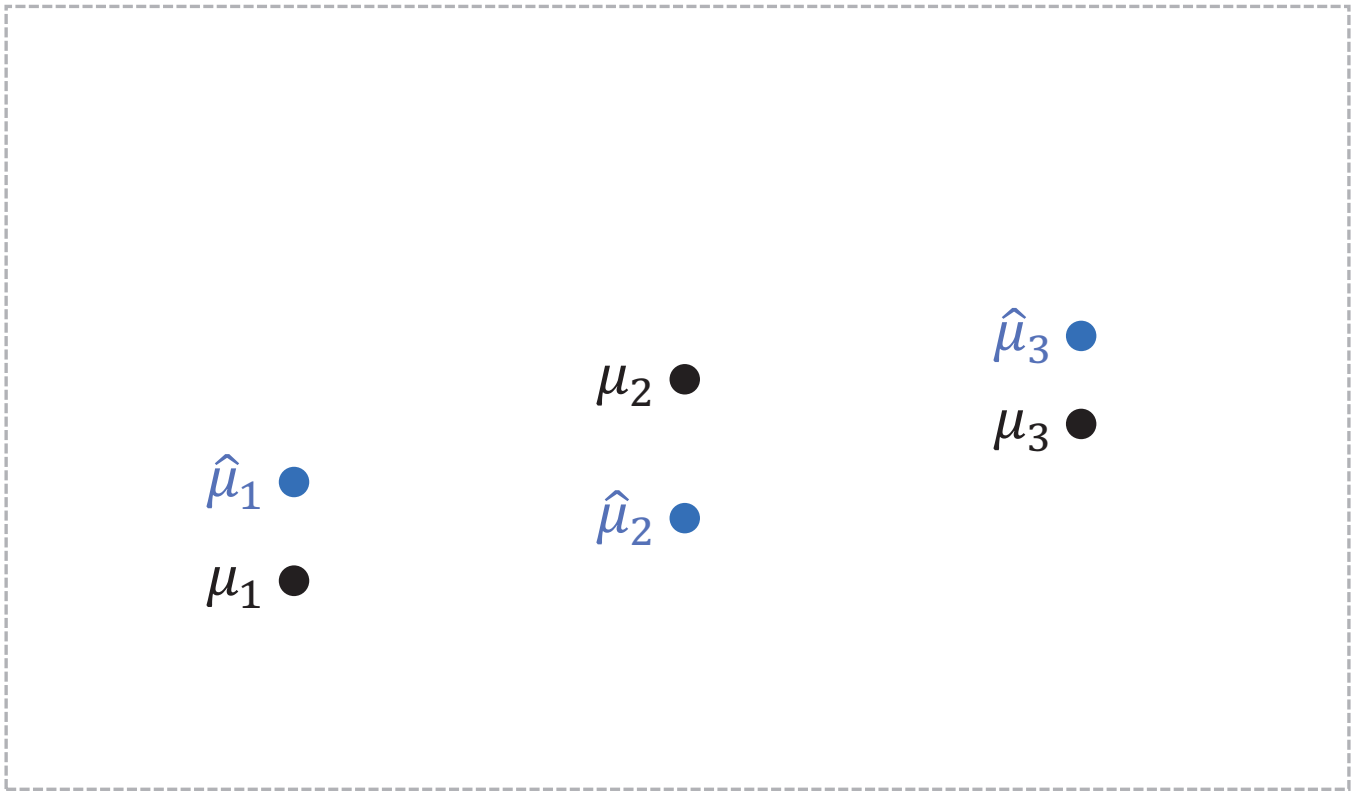


  - A bad regret bound

$$\text{Regret} = T \max_{i \in [K]} \mu_i - \sum_{t=1}^{T} \mu_{i_t} = O(T)$$

Upper Confidence Bound (UCB)

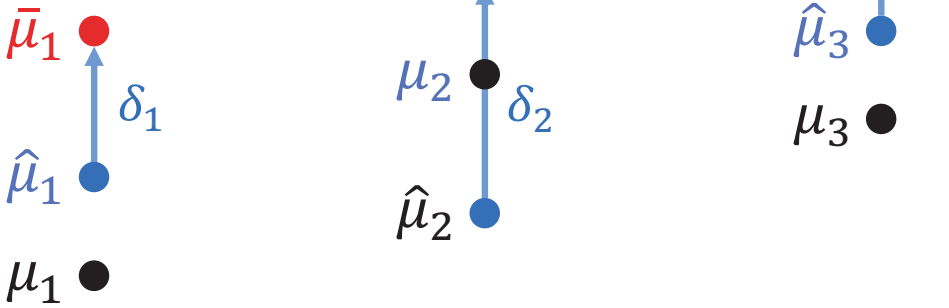
- The Algorithm of UCB



Upper Confidence Bound (UCB)

- The Algorithm of UCB

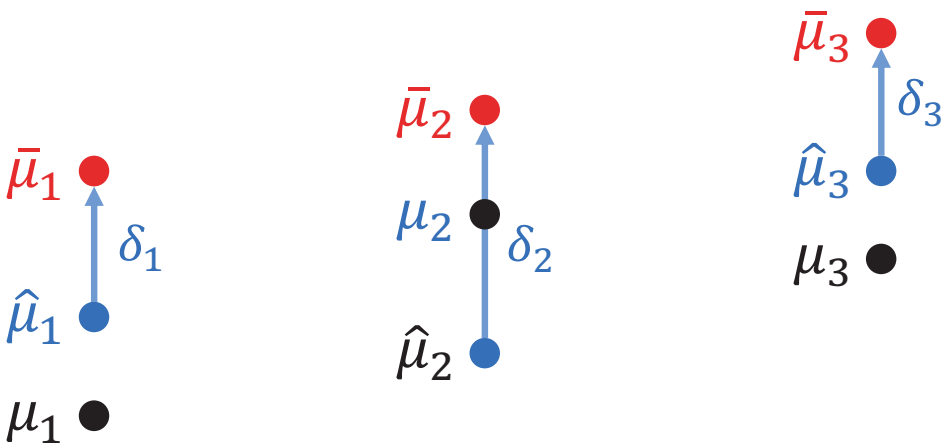
$$\bar{\mu}_i = \hat{\mu}_i + \delta_i$$



Upper Confidence Bound (UCB)

- The Algorithm of UCB

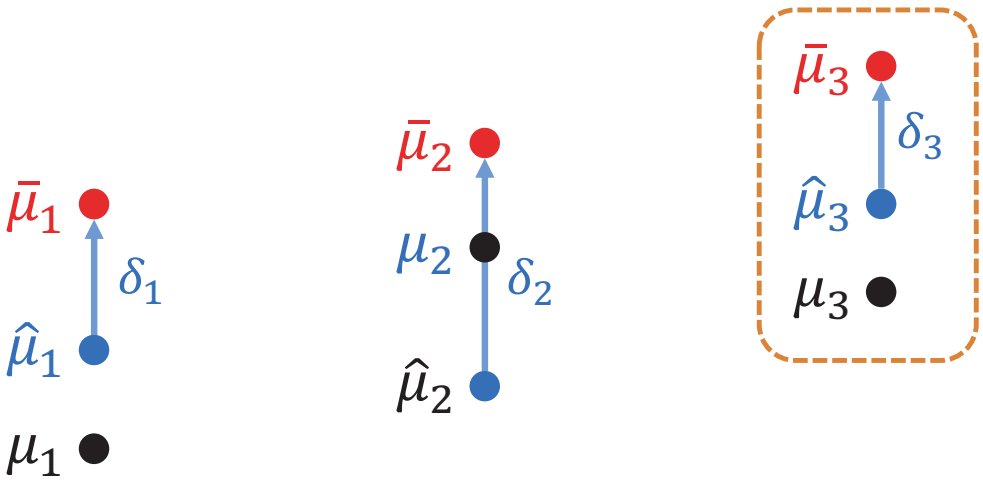
With high probability $\mu_i \leq \bar{\mu}_i = \hat{\mu}_i + \delta_i$



Upper Confidence Bound (UCB)

- The Algorithm of UCB

With high probability $\mu_i \le \bar{\mu}_i = \hat{\mu}_i + \delta_i$

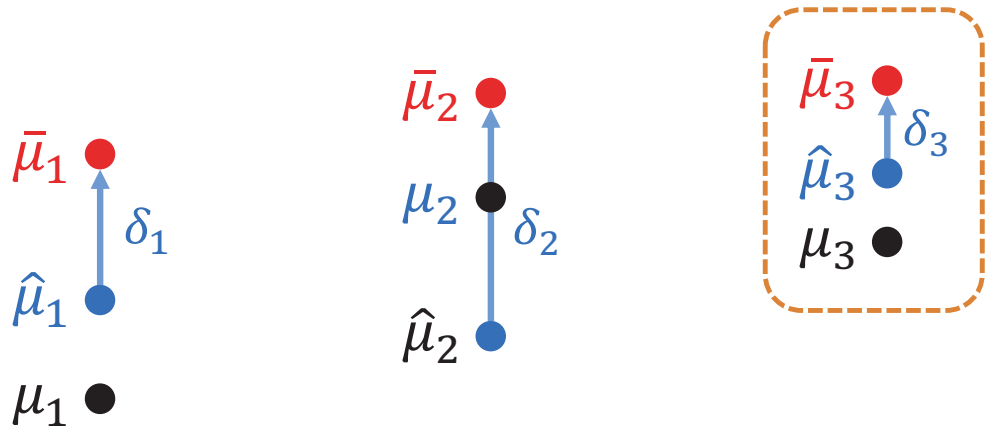


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \bar{\mu}_i$

Upper Confidence Bound (UCB)

- The Algorithm of UCB

With high probability $\mu_i \le \bar{\mu}_i = \hat{\mu}_i + \delta_i$

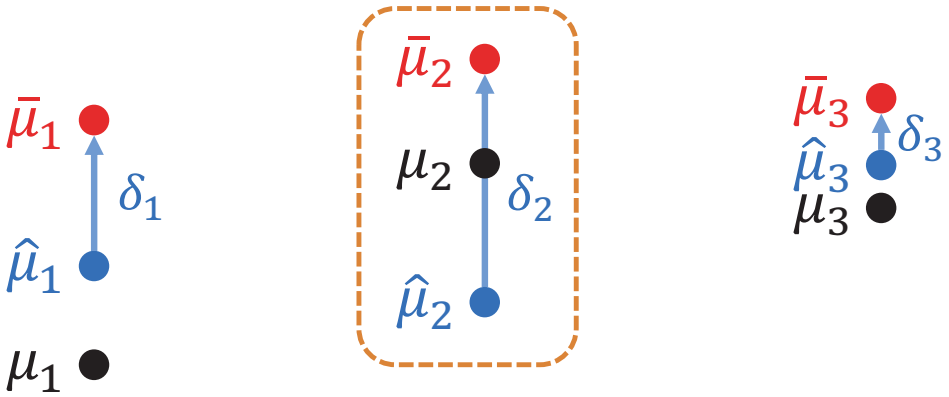


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \bar{\mu}_i$

Upper Confidence Bound (UCB)

The Algorithm of UCB

With high probability $\mu_i \le \bar{\mu}_i = \hat{\mu}_i + \delta_i$

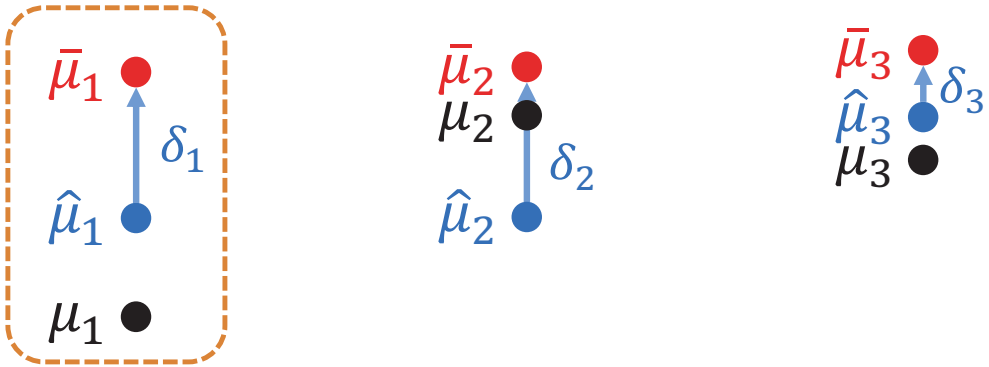


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \bar{\mu}_i$

Upper Confidence Bound (UCB)

- The Algorithm of UCB

With high probability $\mu_i \leq \bar{\mu}_i = \hat{\mu}_i + \delta_i$

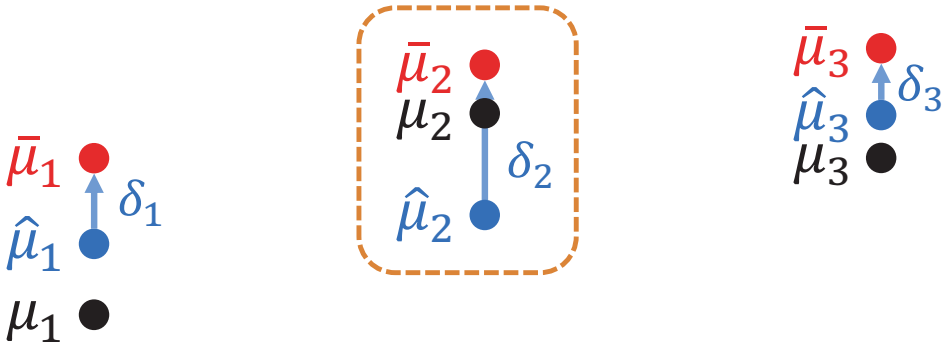


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \bar{\mu}_i$

Upper Confidence Bound (UCB)

- The Algorithm of UCB

With high probability $\mu_i \le \bar{\mu}_i = \hat{\mu}_i + \delta_i$

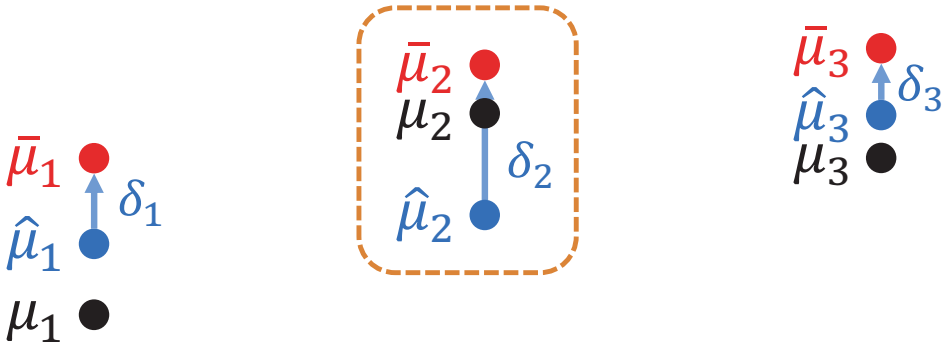


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \bar{\mu}_i$

Upper Confidence Bound (UCB)

The Algorithm of UCB

With high probability $\mu_i \le \bar{\mu}_i = \hat{\mu}_i + \delta_i$



Optimism in Face of Uncertainty: $i_t = \mathrm{argmax}_i \, \bar{\mu}_i$

- Construct $\bar{\mu}_i$ by concentration inequalities (Chernoff-Hoeffding bound) [Auer et al., 2002]

$$\text{Regret} = T \max_{i \in [K]} \mu_i - \sum_{t=1}^{T} \mu_{i_t} \le O(K \log T)$$

LAMDA Learning And Mining from DatA

Outline

LAMDA
Learning And Mining from DatA

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

- Learning with Additional Information
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$

LAMDA Learning And Mining from DatA

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

- Learning with Additional Information
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, the expected reward $\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$

LAMDA Learning And Mining from DatA

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

- Learning with Additional Information
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, the expected reward $\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$

$$\text{Regret} = T \max_{\mathbf{x} \in \mathcal{D}} \mathbf{x}^\top \mathbf{w}_* - \sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{w}_*$$

LAMDA Learning And Mining from DatA

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

- Learning with Additional Information
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, the expected reward $\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$

$$\text{Regret} = T \max_{\mathbf{x} \in \mathcal{D}} \mathbf{x}^\top \mathbf{w}_* - \sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{w}_*$$

Real-valued Feedback [Dani et al., 2008]	
$y = \mathbf{x}^\top \mathbf{w}_* + \epsilon \in \mathbb{R}$	

LAMDA Learning And Mining from DatA

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

- Learning with Additional Information
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, the expected reward $\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$

$$\text{Regret} = T \max_{\mathbf{x} \in \mathcal{D}} \mathbf{x}^\top \mathbf{w}_* - \sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{w}_*$$

Real-valued Feedback [Dani et al., 2008]	One-bit Feedback [Zhang et al., 2016]
$y = \mathbf{x}^\top \mathbf{w}_* + \epsilon \in \mathbb{R}$	Buy or not? Click or not?

LAMDA Learning And Mining from DatA

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

- Learning with Additional Information
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, the expected reward $\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$

$$\text{Regret} = T \max_{\mathbf{x} \in \mathcal{D}} \mathbf{x}^\top \mathbf{w}_* - \sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{w}_*$$

Real-valued Feedback [Dani et al., 2008]	One-bit Feedback [Zhang et al., 2016]
$y = \mathbf{x}^\top \mathbf{w}_* + \epsilon \in \mathbb{R}$	$\Pr[y = \pm 1 \mid \mathbf{x}] = \dfrac{1}{1 + \exp(-y \mathbf{x}^\top \mathbf{w}_*)}$

Linear Optimization under One-bit Feedback

- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K

- Learning with Additional Information
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, the expected reward $\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$

$$\text{Regret} = T \max_{\mathbf{x} \in \mathcal{D}} \mathbf{x}^\top \mathbf{w}_* - \sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{w}_* \leq O(d\sqrt{T})$$

Real-valued Feedback [Dani et al., 2008]	One-bit Feedback [Zhang et al., 2016]
$y = \mathbf{x}^\top \mathbf{w}_* + \epsilon \in \mathbb{R}$	$\Pr[y = \pm 1 \mid \mathbf{x}] = \frac{1}{1 + \exp(-y\mathbf{x}^\top \mathbf{w}_*)}$

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \operatorname{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

Learning History

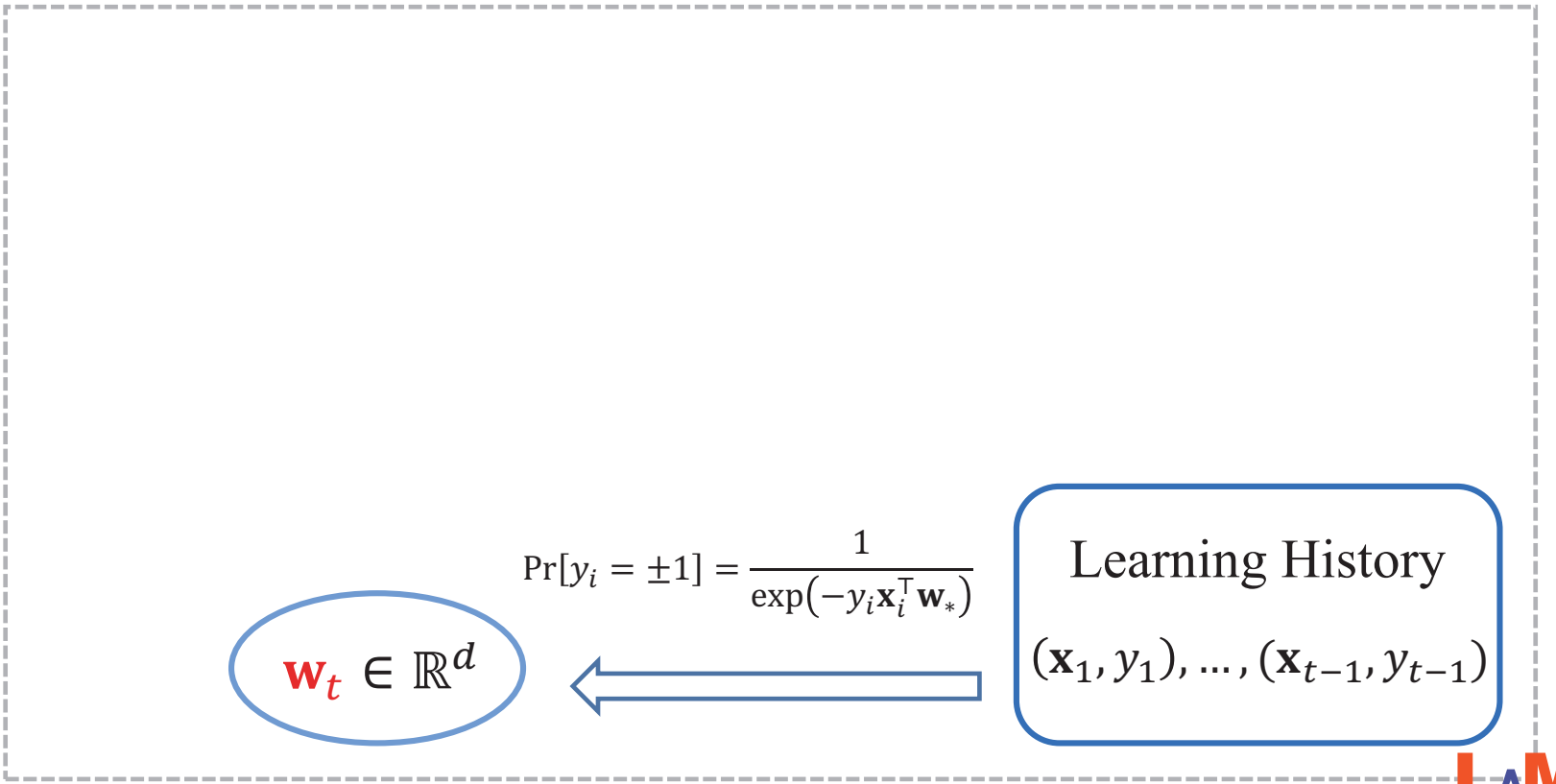
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})$

LAMDA
Learning And Mining from DatA

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$



LAMDA Learning And Mining from DatA

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \operatorname{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

$$\mathbf{w}_* \in \underbrace{\mathcal{C}_t = \text{Ellipse}(\mathbf{w}_t, Z_t, \gamma_t)}_{\text{Confidence Region}}$$

$$\Pr[y_i = \pm 1] = \frac{1}{\exp(-y_i \mathbf{x}_i^\top \mathbf{w}_*)}$$

$\mathbf{w}_t \in \mathbb{R}^d$

Learning History

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})$

LAMDA Learning And Mining from DatA

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \operatorname{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* \leq \underbrace{\max_{\mathbf{w} \in \mathcal{C}_t} \mathbf{x}^\top \mathbf{w}}_{\text{Upper Bound}} = \bar{\mu}_{\mathbf{x}}$$

$$\mathbf{w}_* \in \underbrace{\mathcal{C}_t = \text{Ellipse}(\mathbf{w}_t, Z_t, \gamma_t)}_{\text{Confidence Region}}$$

$$\mathbf{w}_t \in \mathbb{R}^d$$

$$\Pr[y_i = \pm 1] = \frac{1}{\exp(-y_i \mathbf{x}_i^\top \mathbf{w}_*)}$$

Learning History

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})$$

LAMDA Learning And Mining from DatA

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* \leq \underbrace{\max_{\mathbf{w} \in \mathcal{C}_t} \mathbf{x}^\top \mathbf{w} = \bar{\mu}_{\mathbf{x}}}_{\text{Upper Bound}} \Longrightarrow \mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$$

$$\mathbf{w}_* \in \underbrace{\mathcal{C}_t = \text{Ellipse}(\mathbf{w}_t, Z_t, \gamma_t)}_{\text{Confidence Region}}$$

$$\mathbf{w}_t \in \mathbb{R}^d \Longleftarrow \Pr[y_i = \pm 1] = \frac{1}{\exp(-y_i \mathbf{x}_i^\top \mathbf{w}_*)} \Longleftarrow \text{Learning History } (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})$$

LAMDA Learning And Mining from DatA

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* \leq \underbrace{\max_{\mathbf{w} \in \mathcal{C}_t} \mathbf{x}^\top \mathbf{w}}_{\text{Upper Bound}} = \bar{\mu}_{\mathbf{x}} \Longrightarrow \mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$$

Submit $\mathbf{x}_t$

Observe $y_t \in \{\pm 1\}$

$$\mathbf{w}_* \in \underbrace{\mathcal{C}_t = \text{Ellipse}(\mathbf{w}_t, Z_t, \gamma_t)}_{\text{Confidence Region}}$$

$$\Pr[y_i = \pm 1] = \frac{1}{\exp(-y_i \mathbf{x}_i^\top \mathbf{w}_*)}$$

Learning History

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})$

$\mathbf{w}_t \in \mathbb{R}^d$

LAMDA Learning And Mining from DatA

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* \leq \underbrace{\max_{\mathbf{w} \in \mathcal{C}_t} \mathbf{x}^\top \mathbf{w}}_{\text{Upper Bound}} = \bar{\mu}_{\mathbf{x}}$ ⟹ $\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$

Submit $\mathbf{x}_t$

Observe $y_t \in \{\pm 1\}$

Save $(\mathbf{x}_t, y_t)$

Learning History $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})$

$\Pr[y_i = \pm 1] = \frac{1}{\exp(-y_i \mathbf{x}_i^\top \mathbf{w}_*)}$

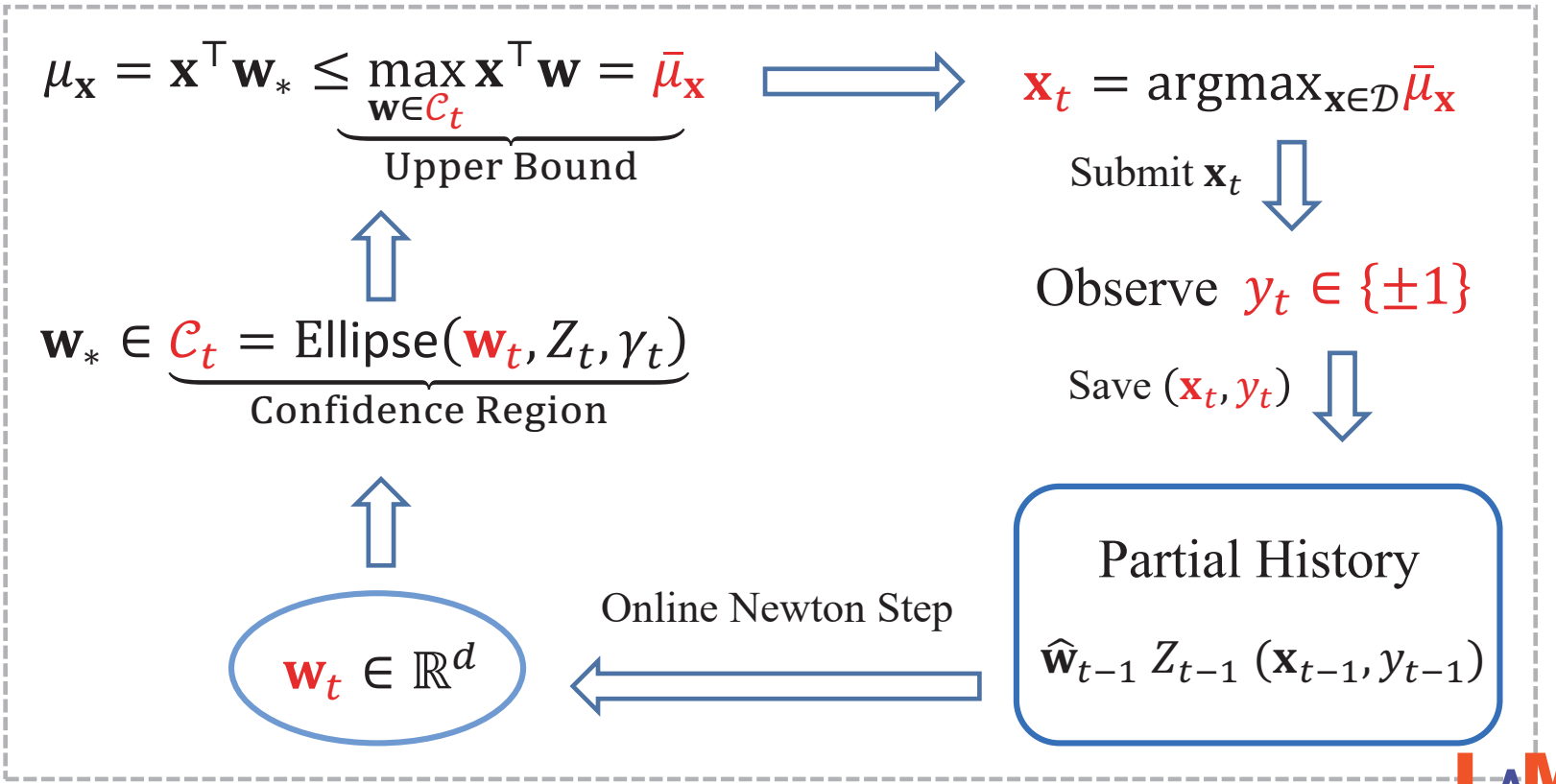
$\mathbf{w}_t \in \mathbb{R}^d$

$\mathbf{w}_* \in \underbrace{\mathcal{C}_t = \text{Ellipse}(\mathbf{w}_t, Z_t, \gamma_t)}_{\text{Confidence Region}}$

A UCB-Type Algorithm [Zhang et al., 2016]

In the t-th Round

- Construct an upper bound $\bar{\mu}_{\mathbf{x}} \geq \mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_*$ for each arm $\mathbf{x} \in \mathcal{D}$
- By Optimism in Face of Uncertainty, $\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \bar{\mu}_{\mathbf{x}}$



LAMDA Learning And Mining from DatA

Implementation Issues [Zhang et al., 2016]

- Online Newton Step
  - Given $\mathbf{w}_{t-1}$ and $(\mathbf{x}_{t-1}, y_{t-1})$
  - Define $f_{t-1}(\mathbf{w}) = \log(1 + \exp(-y_{t-1}\mathbf{x}_{t-1}^\top \mathbf{w}))$

$$\mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w}} \frac{\|\mathbf{w} - \mathbf{w}_t\|_{Z_t}^2}{2} + (\mathbf{w} - \mathbf{w}_{t-1})^\top \nabla f_{t-1}(\mathbf{w}_{t-1})$$

where $Z_t = Z_{t-1} + \frac{\beta}{2}\mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top$

- Arm Selection

$$\mathbf{x}_t = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}} \underbrace{\max_{\mathbf{w}\in\mathcal{C}_t} \mathbf{x}^\top \mathbf{w}}_{:=\bar{\mu}_{\mathbf{x}}} = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}} \max_{\|\mathbf{w}-\mathbf{w}_t\|_{Z_t} \le \sqrt{\gamma_t}} \mathbf{x}^\top \mathbf{w}$$

  - The above problem is NP-hard in general
  - Tractable when $\mathcal{D}$ is discrete or a ball

LAMDA Learning And Mining from DatA

Outline

Regret Bound [Zhang et al., 2016]

Theorem 1 (Confidence Region)

With a high probability, we have

$$(\mathbf{w}_* - \mathbf{w}_t)^\top \left(\lambda I + \frac{\beta}{2} \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i^\top \right) (\mathbf{w}_* - \mathbf{w}_t) \leq \gamma_t = O(d \log t)$$

for all $t > 0$.

- Optimality condition of online Newton step
- Bernstein's inequality for martingales
- Peeling technique for decoupling the dependence

LAMDA Learning And Mining from DatA

Regret Bound [Zhang et al., 2016]

Theorem 1 (Confidence Region)

With a high probability, we have

$$(\mathbf{w}_* - \mathbf{w}_t)^\top \left(\lambda I + \frac{\beta}{2} \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i^\top \right) (\mathbf{w}_* - \mathbf{w}_t) \leq \gamma_t = O\left(d \log t\right)$$

for all $t > 0$.

- Optimality condition of online Newton step
- Bernstein's inequality for martingales
- Peeling technique for decoupling the dependence

Theorem 2 (Regret)

With a high probability, we have

$$T \max_{\mathbf{x} \in \mathcal{D}} \mathbf{x}^\top \mathbf{w}_* - \sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{w}_* \leq 4 \sqrt{\frac{\gamma_T T}{\beta} \log \frac{\det(Z_{T+1})}{\det(Z_1)}} = O\left(d\sqrt{T}\right)$$

for all $T > 0$.

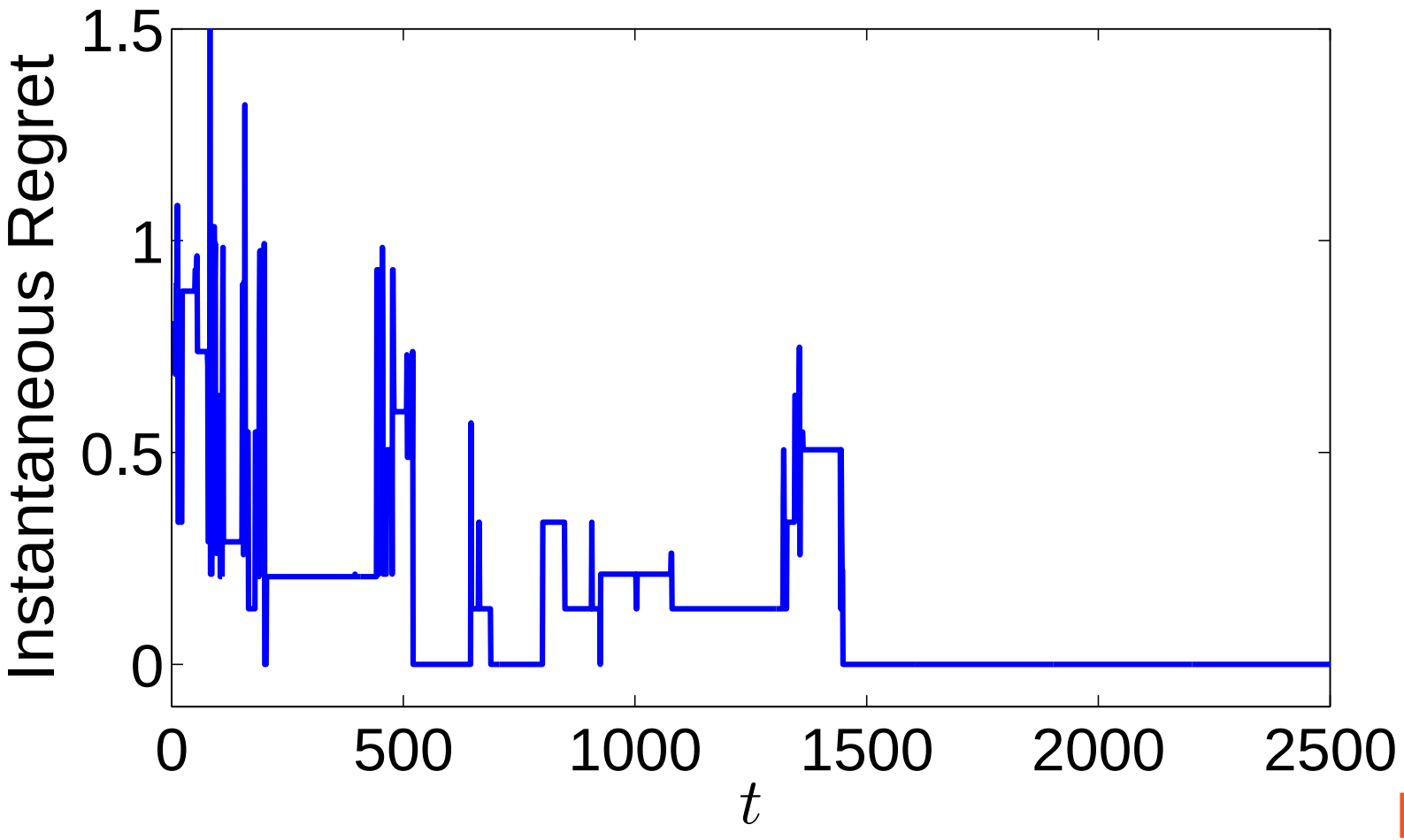
LAMDA Learning And Mining from DatA

Outline

LAMDA Learning And Mining from DatA

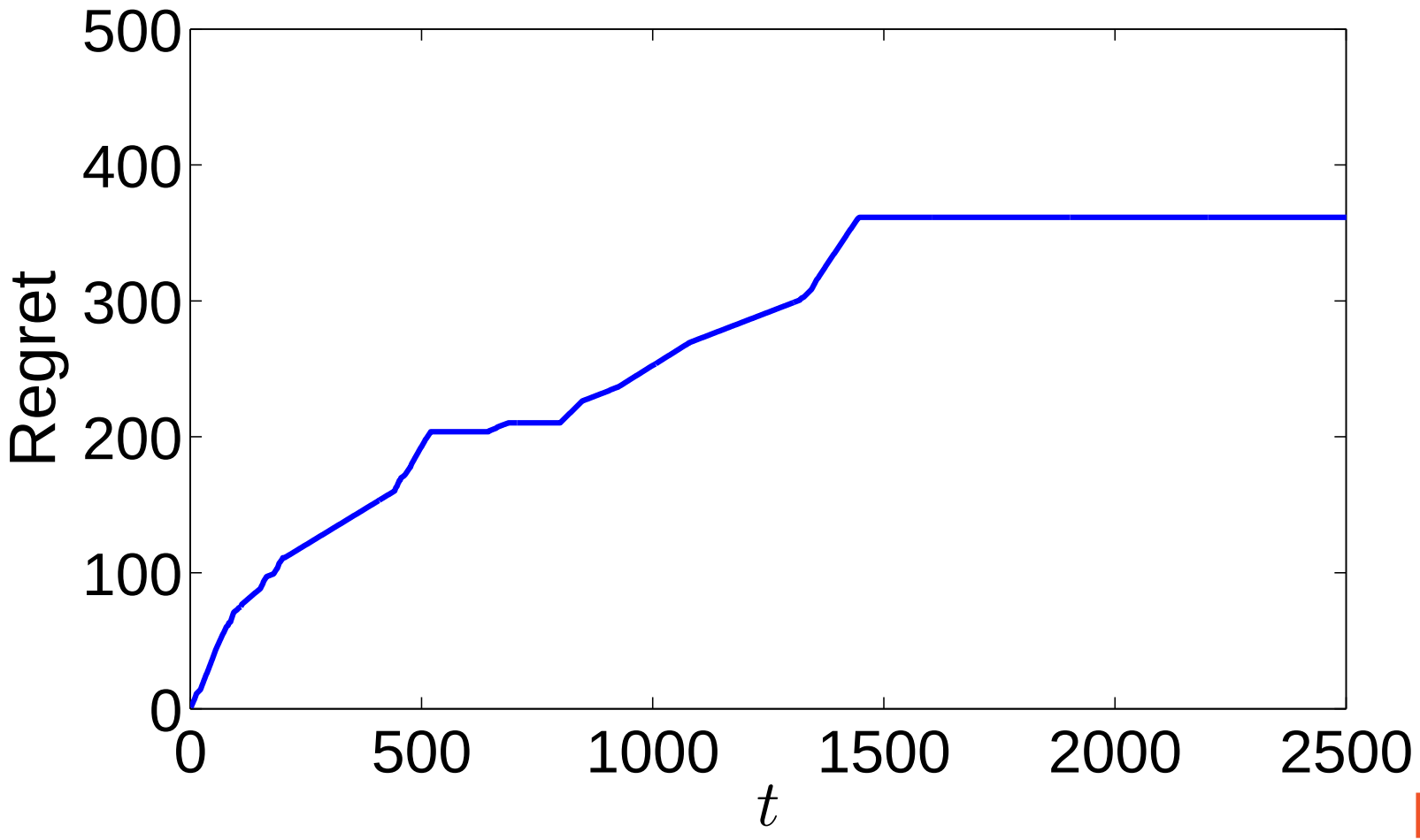
Experimental Results

- $\mathcal{D} \subseteq \mathbb{R}^{10}$ and $|\mathcal{D}| = 100$



Experimental Results

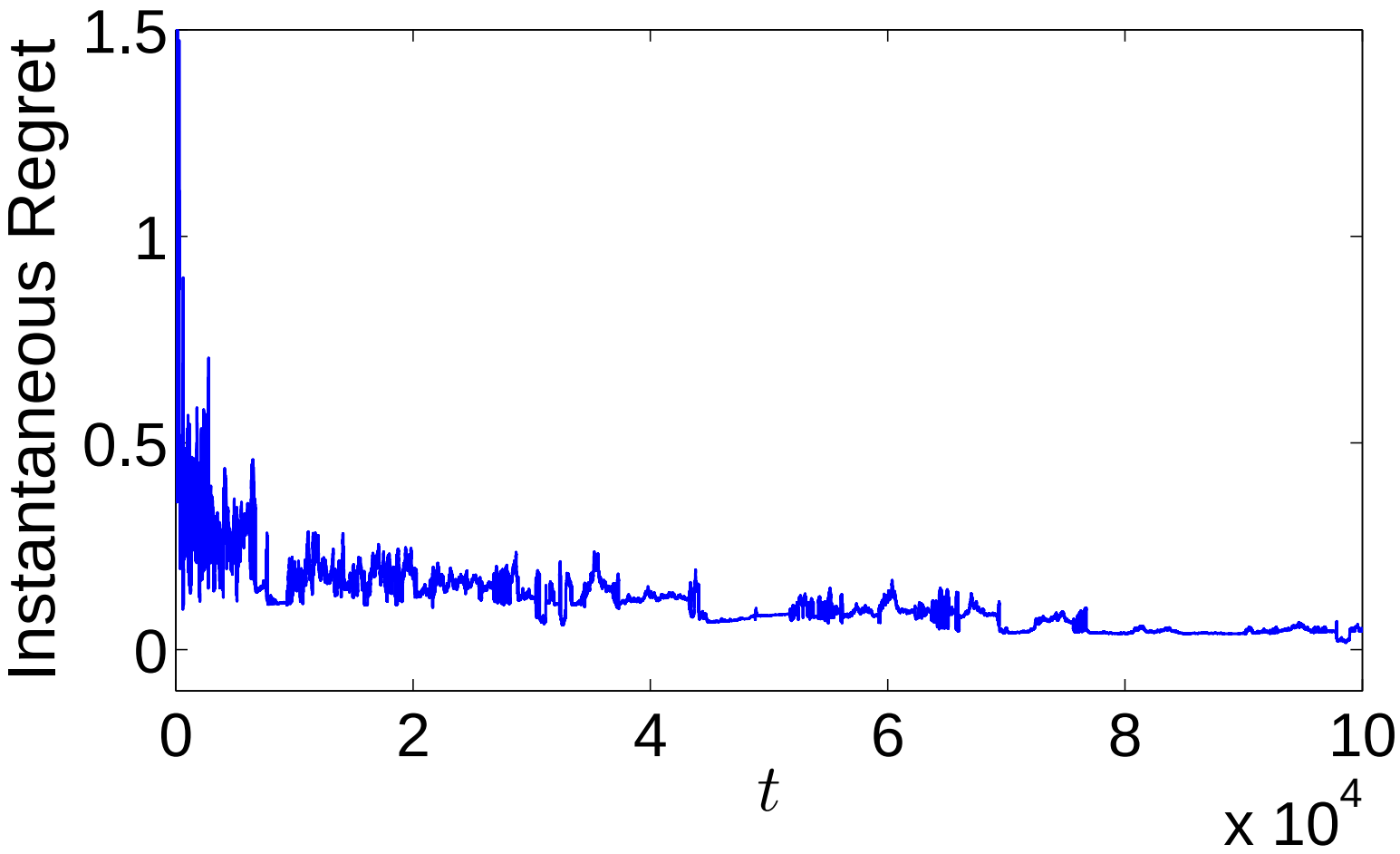
- $\mathcal{D} \subseteq \mathbb{R}^{10}$ and $|\mathcal{D}| = 100$



LAMDA
Learning And Mining from DatA

Experimental Results

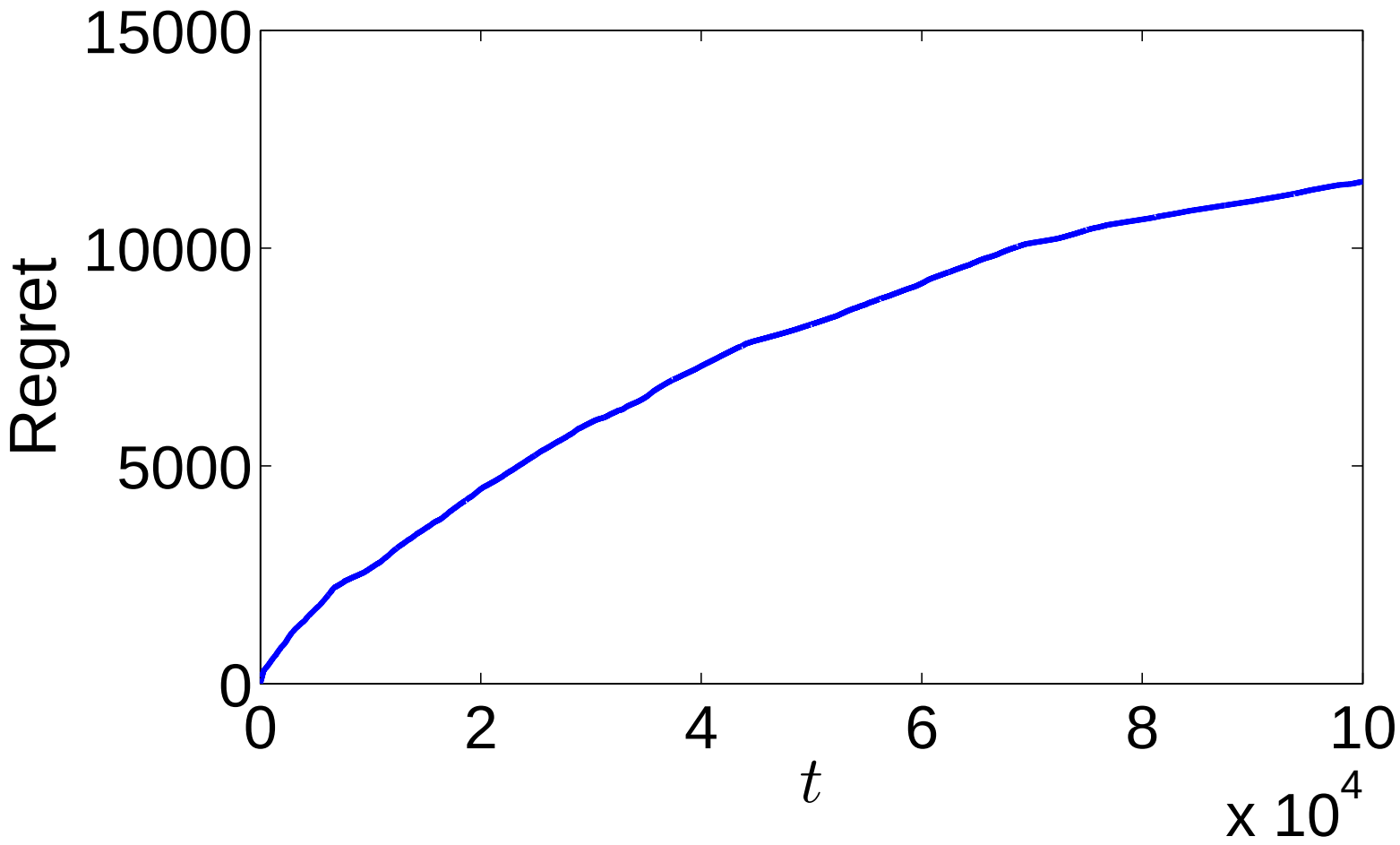
- $\mathcal{D} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\} \subseteq \mathbb{R}^{10}$ and $|\mathcal{D}| = \infty$



LAMDA
Learning And Mining from DatA

Experimental Results

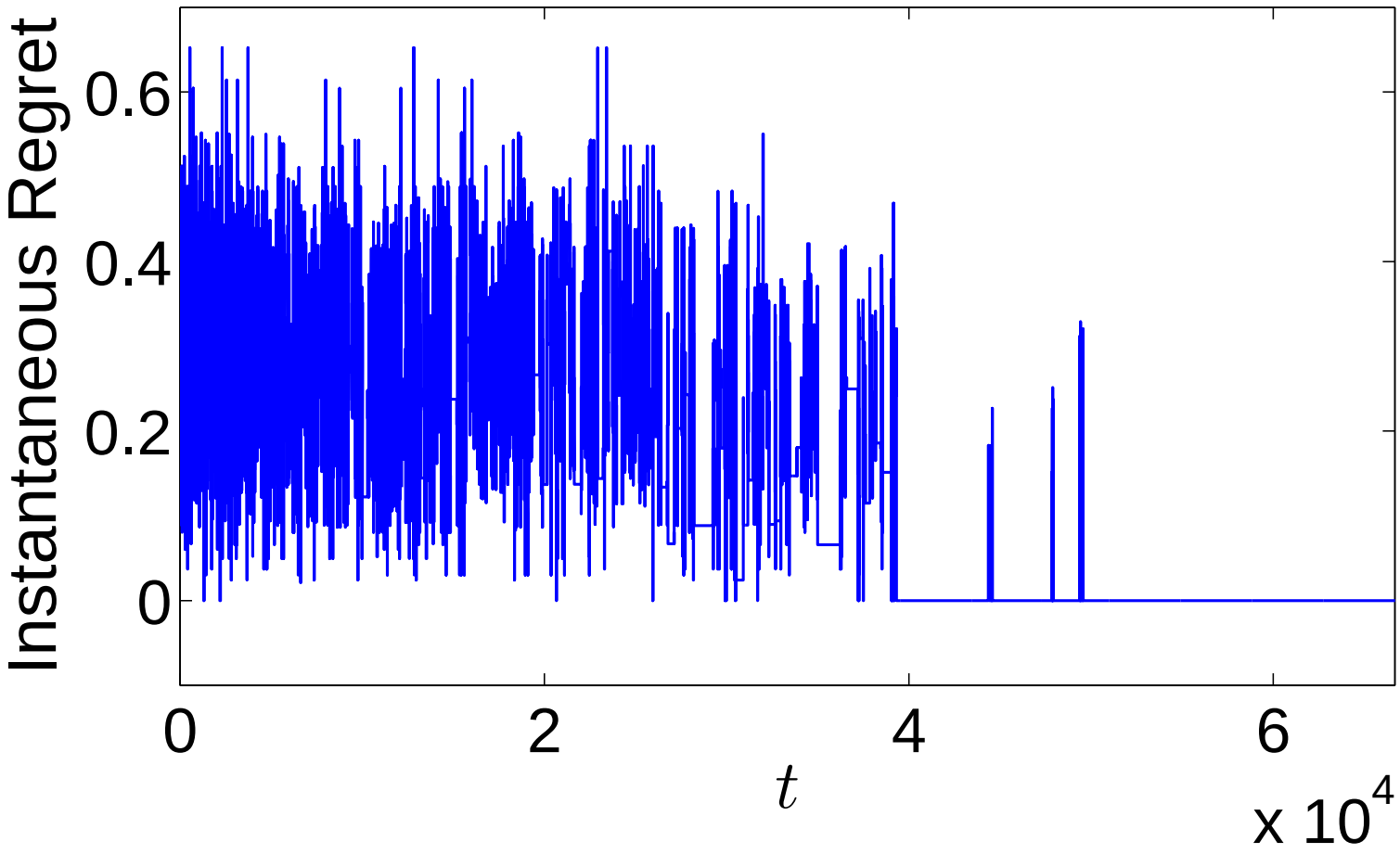
- $\mathcal{D} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\} \subseteq \mathbb{R}^{10}$ and $|\mathcal{D}| = \infty$



LAMDA
Learning And Mining from DatA

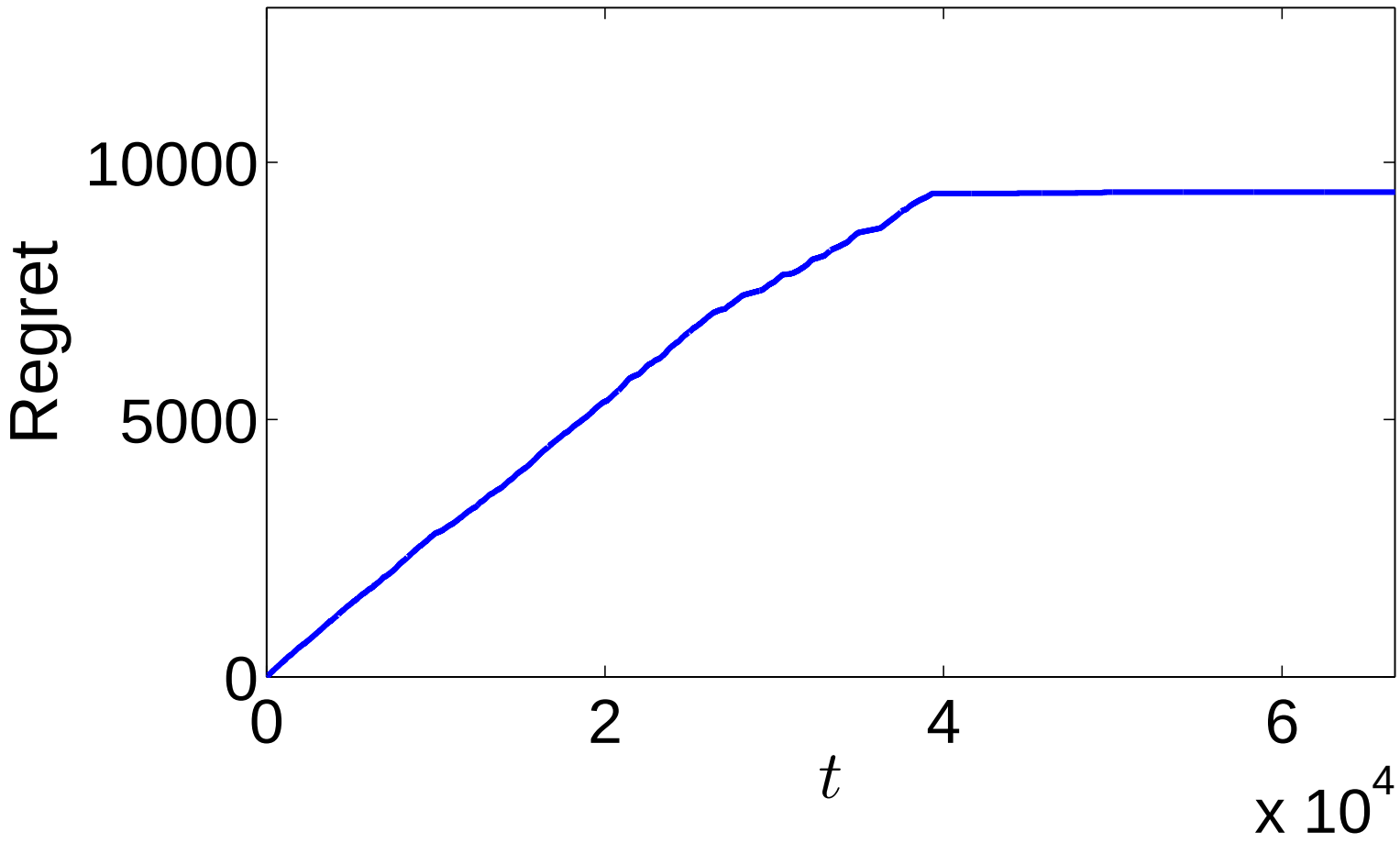
Experimental Results

- $\mathcal{D} \subseteq \mathbb{R}^{100}$ and $|\mathcal{D}| = 1000$



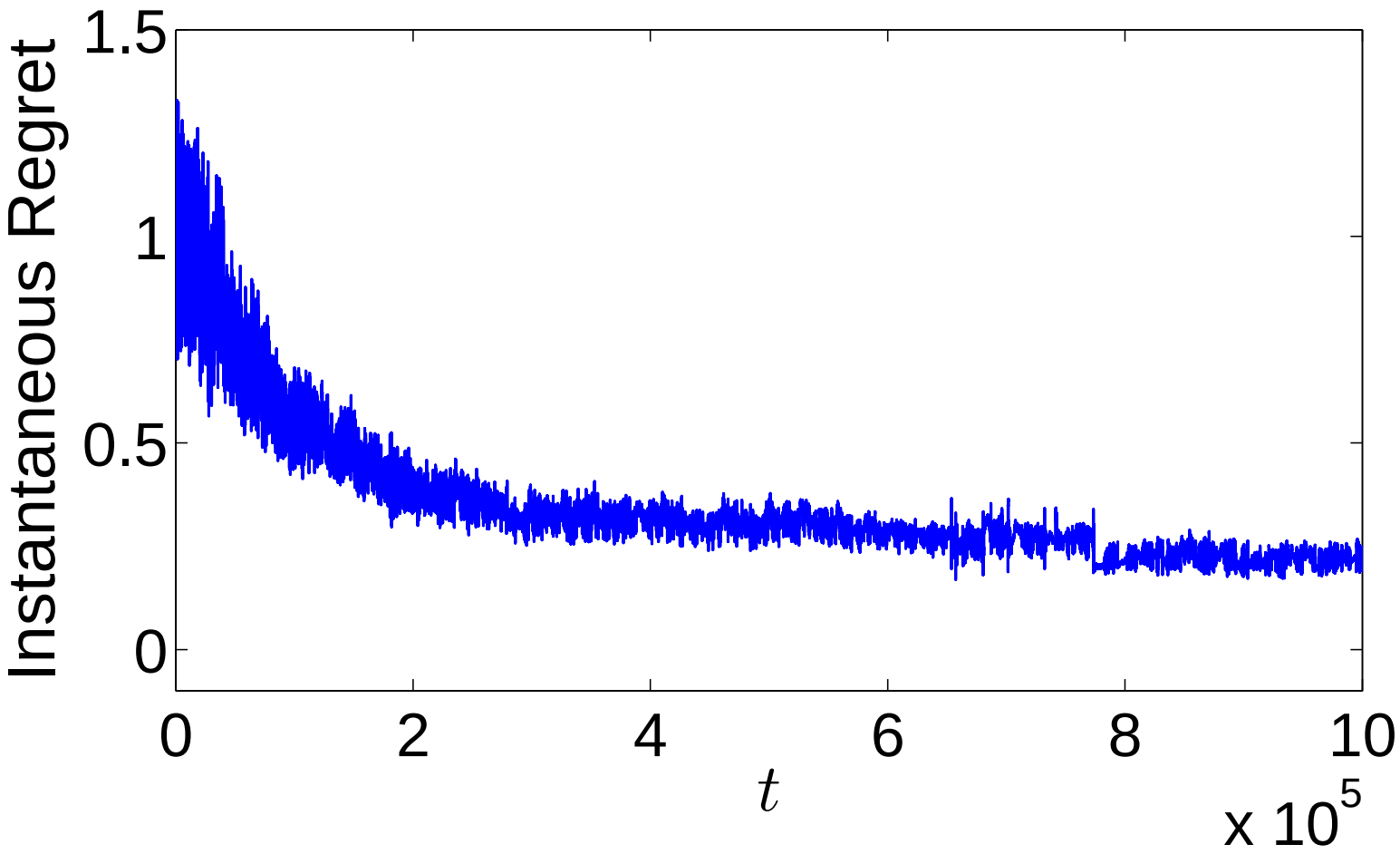
Experimental Results

- $\mathcal{D} \subseteq \mathbb{R}^{100}$ and $|\mathcal{D}| = 1000$



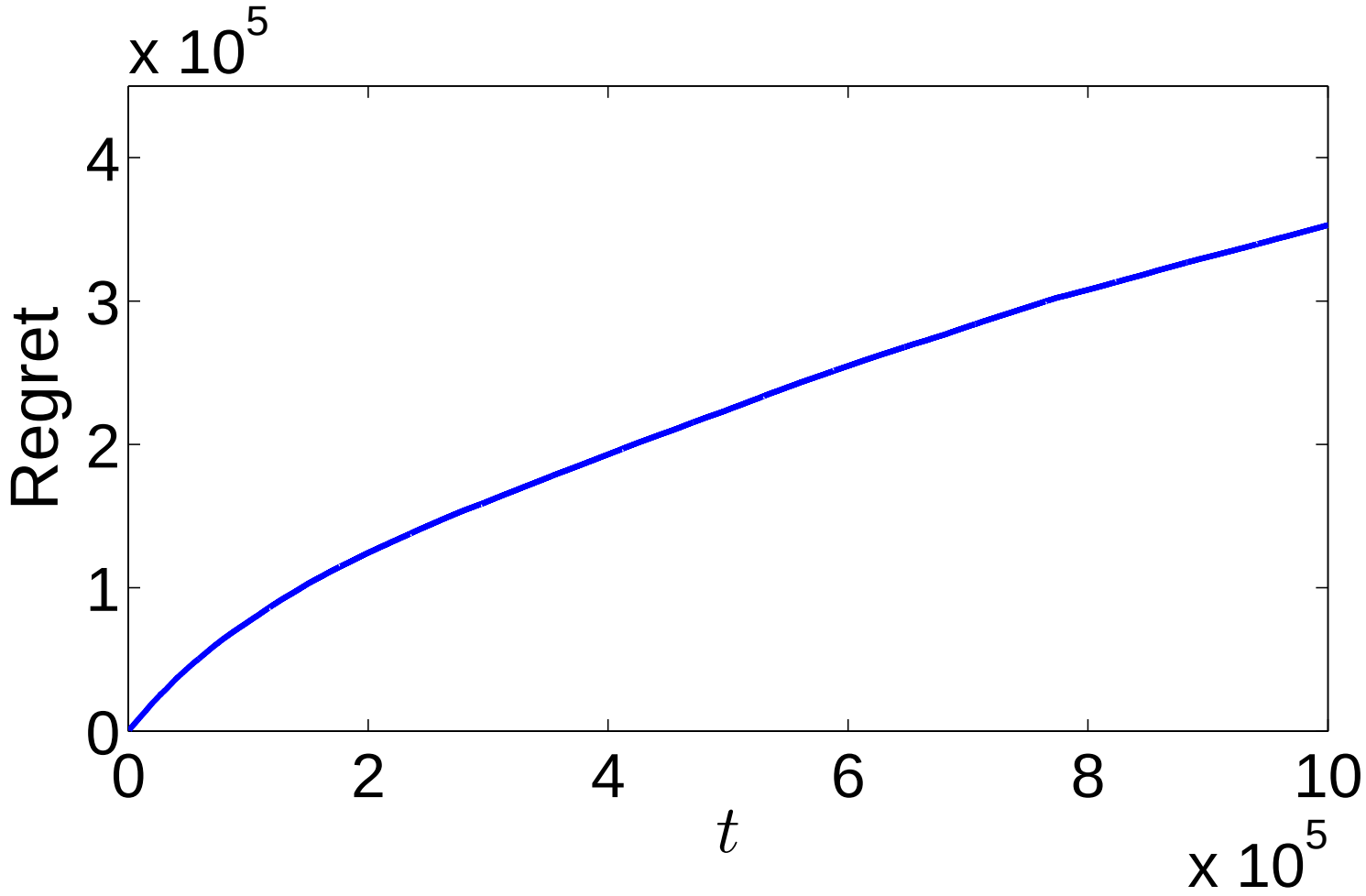
Experimental Results

- $\mathcal{D} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\} \subseteq \mathbb{R}^{100}$ and $|\mathcal{D}| = \infty$



Experimental Results

- $\mathcal{D} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\} \subseteq \mathbb{R}^{100}$ and $|\mathcal{D}| = \infty$

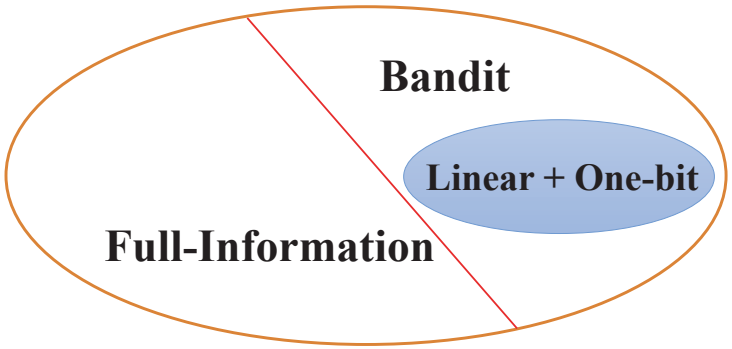


LAMDA
Learning And Mining from DatA

Conclusion and Future Work

Conclusion

- Online stochastic linear optimization under one-bit feedback
- An efficient algorithm with $O(d\sqrt{T})$ regret bound



Future Work

- Parameter selection in practice
- More observation models, such as the $\mathrm{sign}(\cdot)$
- Select multiple arms in each round
- $\mathbf{w}_*$ may change from round to round

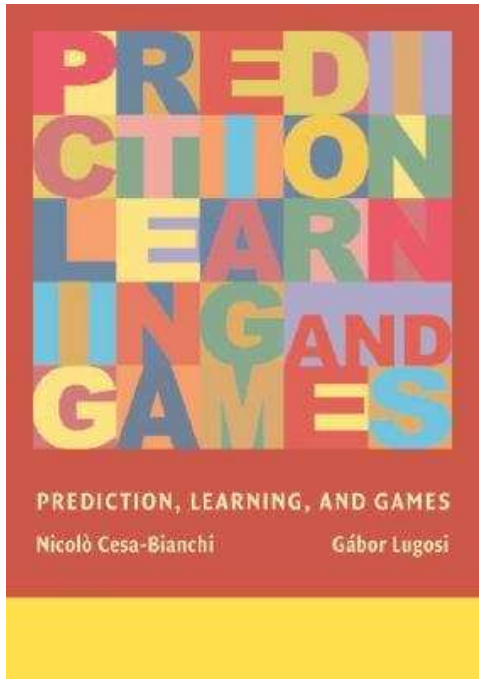
LAMDA
Learning And Mining from DatA

Resources

- Book
  - Prediction, Learning, and Games [Cesa-Bianchi and Lugosi, 2006]

(a) Cesa-Bianchi

(b) Lugosi



- Surveys
  - Online Learning and Online Convex Optimization [Shalev-Shwartz, 2011]
  - Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems [Bubeck and Cesa-Bianchi, 2012]

LAMDA Learning And Mining from DatA

Reference I

Thanks!

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002).

Finite-time analysis of the multiarmed bandit problem.
Machine Learning, 47(2-3):235–256.

Bubeck, S. and Cesa-Bianchi, N. (2012).

Regret analysis of stochastic and nonstochastic multi-armed bandit problems.
Foundations and Trends in Machine Learning, 5(1):1–122.

Cesa-Bianchi, N. and Lugosi, G. (2006).

Prediction, Learning, and Games.
Cambridge University Press.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008).

Stochastic linear optimization under bandit feedback.
In *Proceedings of the 21st Annual Conference on Learning*, pages 355–366.

Karp, R. M. (1992).

On-line algorithms versus off-line algorithms: How much is it worth to know the future?
In *Proceedings of the IFIP 12th World Computer Congress on Algorithms, Software, Architecture – Information Processing*, pages 416–429.

Robbins, H. (1952).

Some aspects of the sequential design of experiments.
Bulletin of the American Mathematical Society, 58(5):527–535.

LAMDA
Learning And Mining from DatA

Reference II

Shalev-Shwartz, S. (2011).

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

Zhang, L., Jin, R., Chen, C., Bu, J., and He, X. (2012).

Efficient online learning for large-scale sparse kernel logistic regression.
In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1219–1225.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-H. (2016).

Online stochastic linear optimization under one-bit feedback.
In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*.

Zhang, L., Yang, T., Jin, R., and Zhou, Z.-H. (2015).

Online bandit learning for a special class of non-convex losses.
In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 3158–3164.

Zhang, L., Yi, J., Jin, R., Lin, M., and He, X. (2013).

Online kernel learning with a near optimal sparsity bound.
In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

LAMDA
Learning And Mining from DatA