Bandit Online Learning

Lijun Zhang

Nanjing University, China

June 23, 2017

Outline

Outline

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

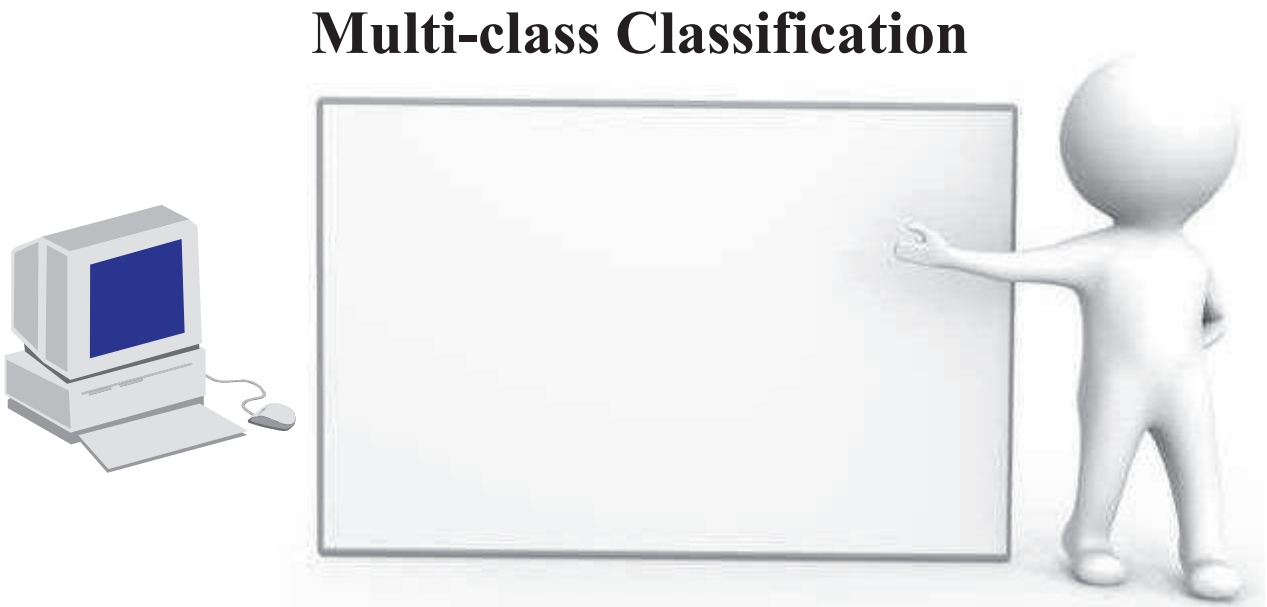
- Bandit Online Learning

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



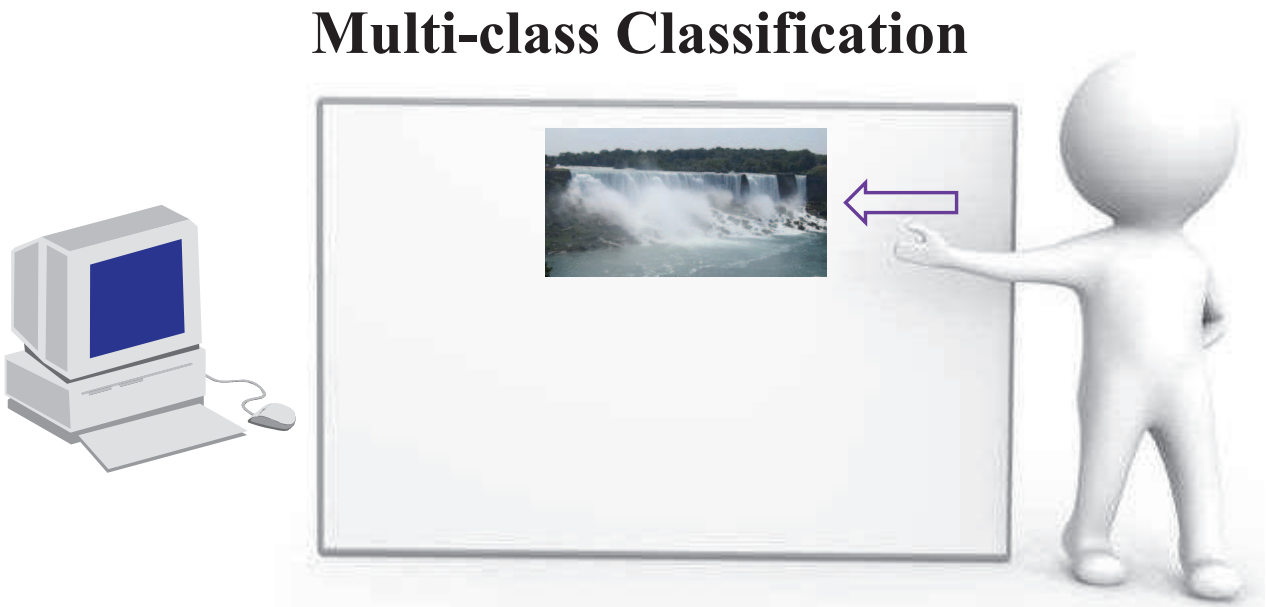
- Bandit Online Learning

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



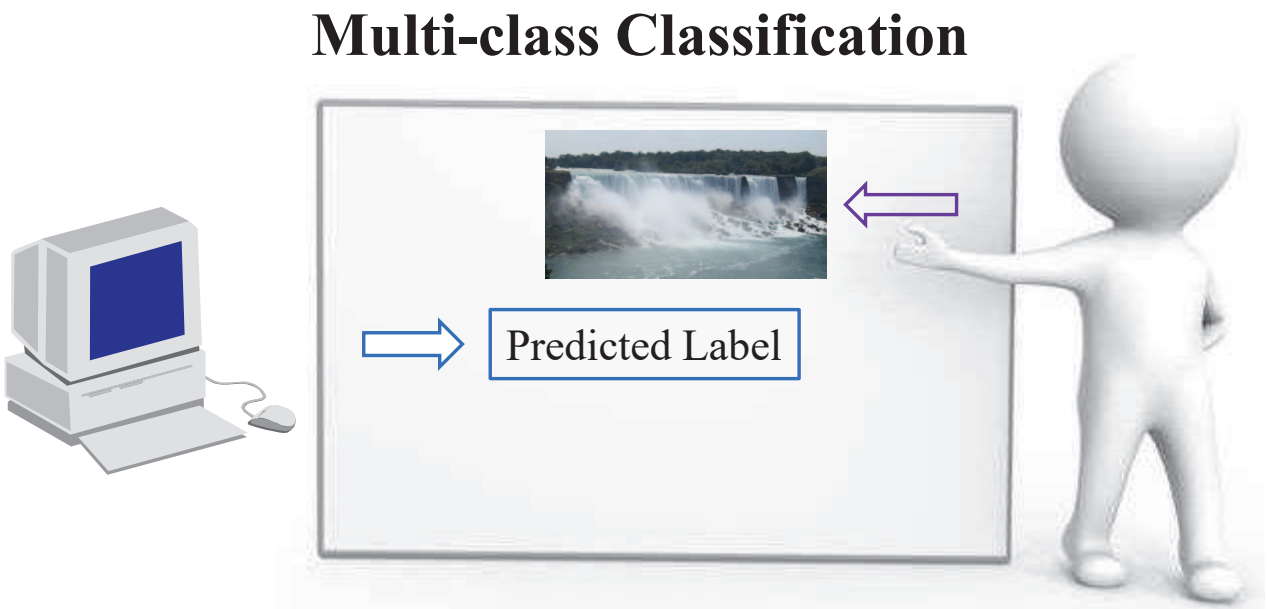
- Bandit Online Learning

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



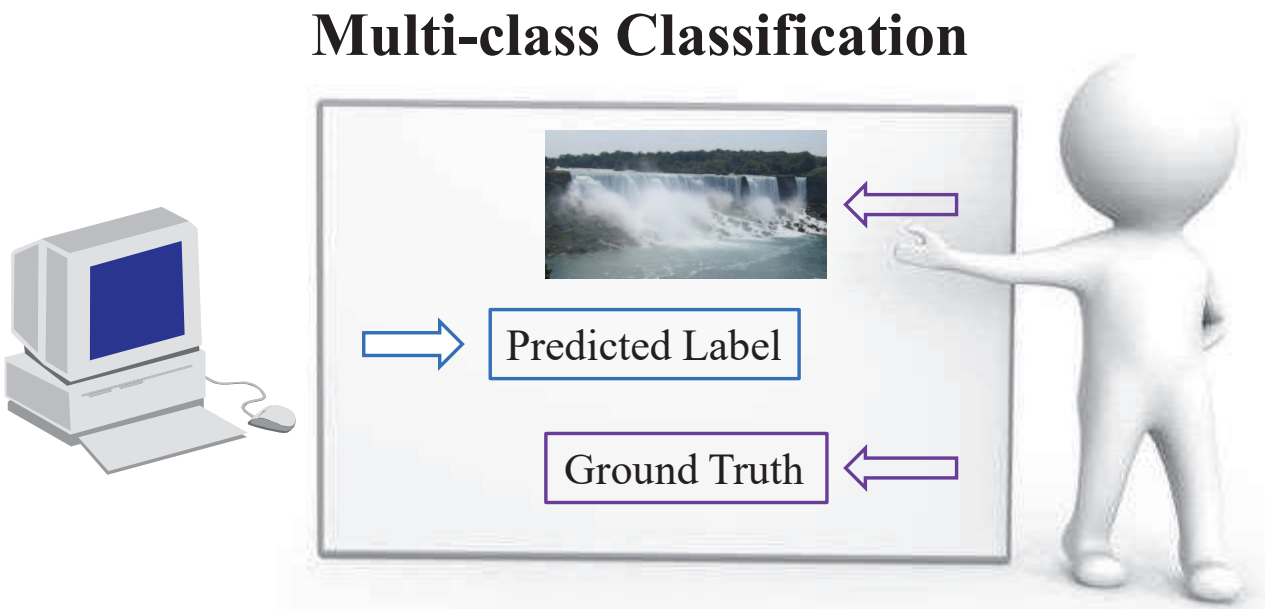
- Bandit Online Learning

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



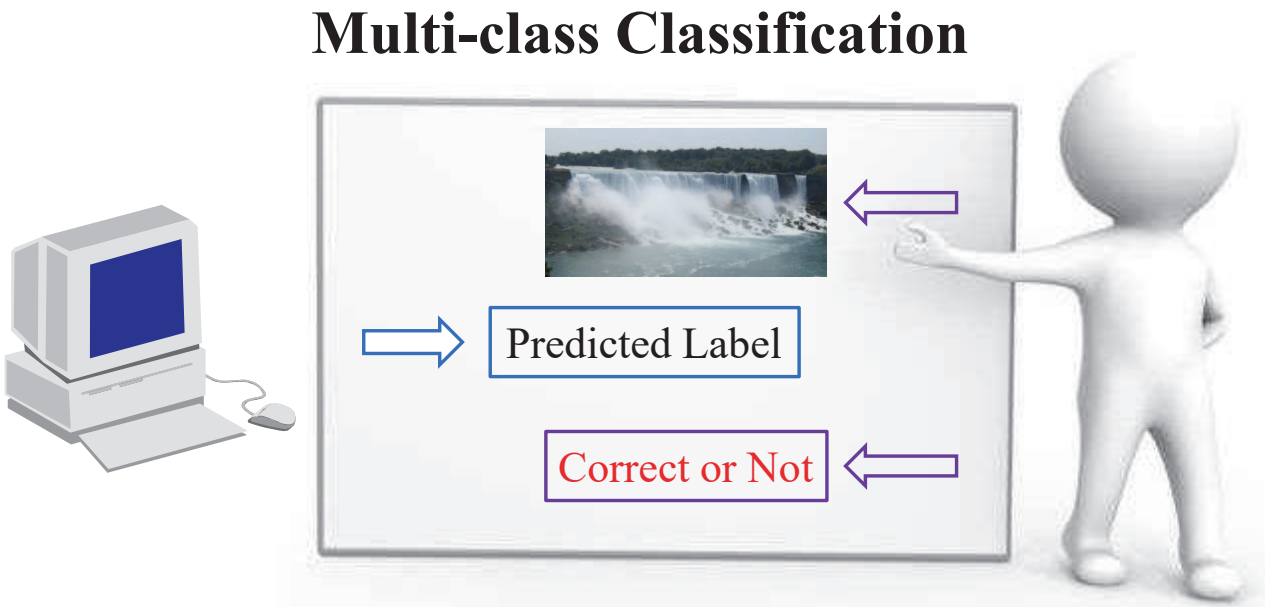
- Bandit Online Learning

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**
2: Learner picks a decision $\mathbf{x}_t \in \mathcal{D}$
Adversary chooses a function $f_t(\cdot)$
3: Learner *only* observes loss $f_t(\mathbf{x}_t)$ and updates $\mathbf{x}_t$
4: **end for**

Cumulative Loss

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

Regret

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

LAMDA
Learning And Mining from DatA

Statistical Assumptions

- Stochastic
  - $f_1, \ldots, f_t$ are i.i.d.

Statistical Assumptions

- Stochastic
  - $f_1, \ldots, f_t$ are i.i.d.
- Adversarial
  - Oblivious: f_t is independent of $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$

Examination

  - Nonoblivious: f_t depends on $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$

Interview

LAMDA Learning And Mining from DatA

Outline

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

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

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	$X_{3,1}$

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

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

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	$X_{3,3}$	$X_{3,4}$

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	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$

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

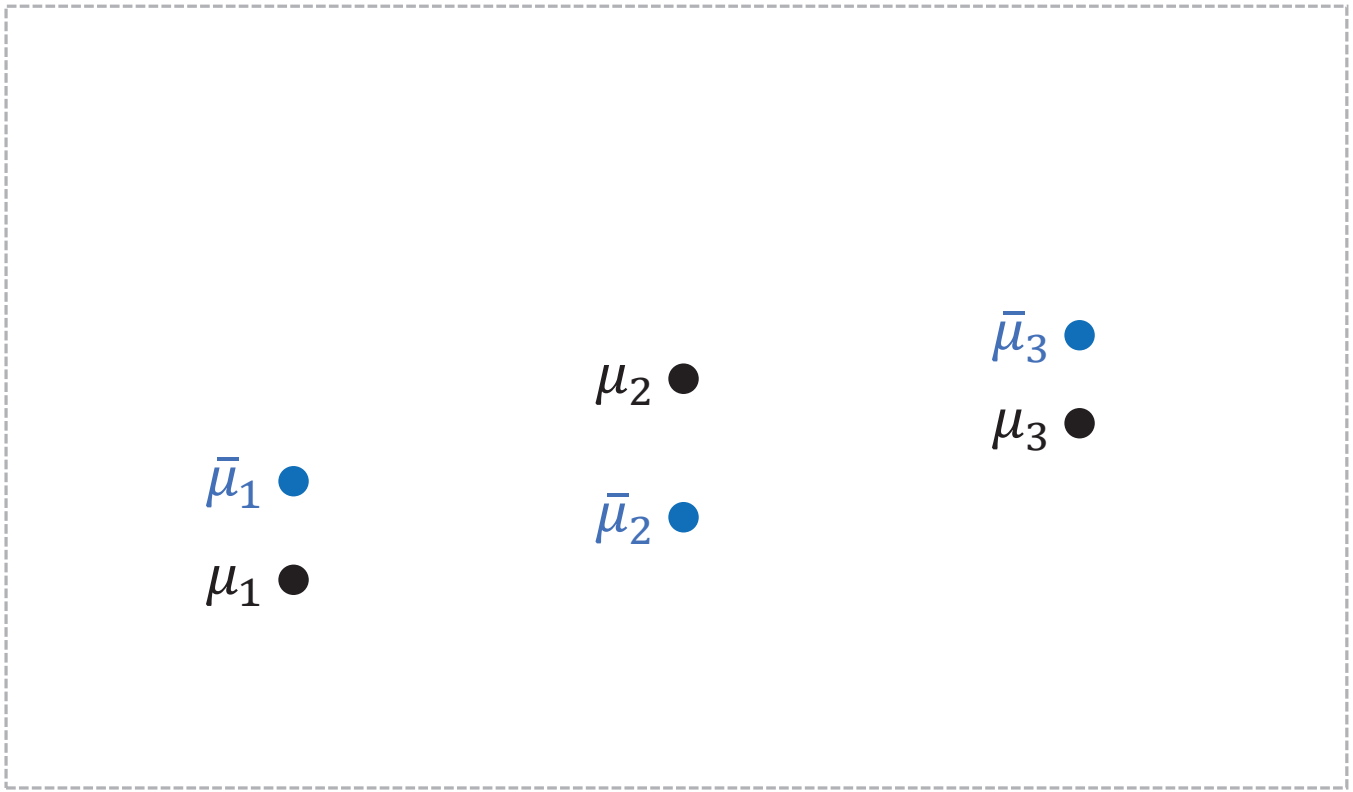
$$\text{Regret} = T \max_{i \in [K]} \mu_i - \sum_{t=1}^{T} \mu_{i_t}$$

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

LAMDA Learning And Mining from DatA

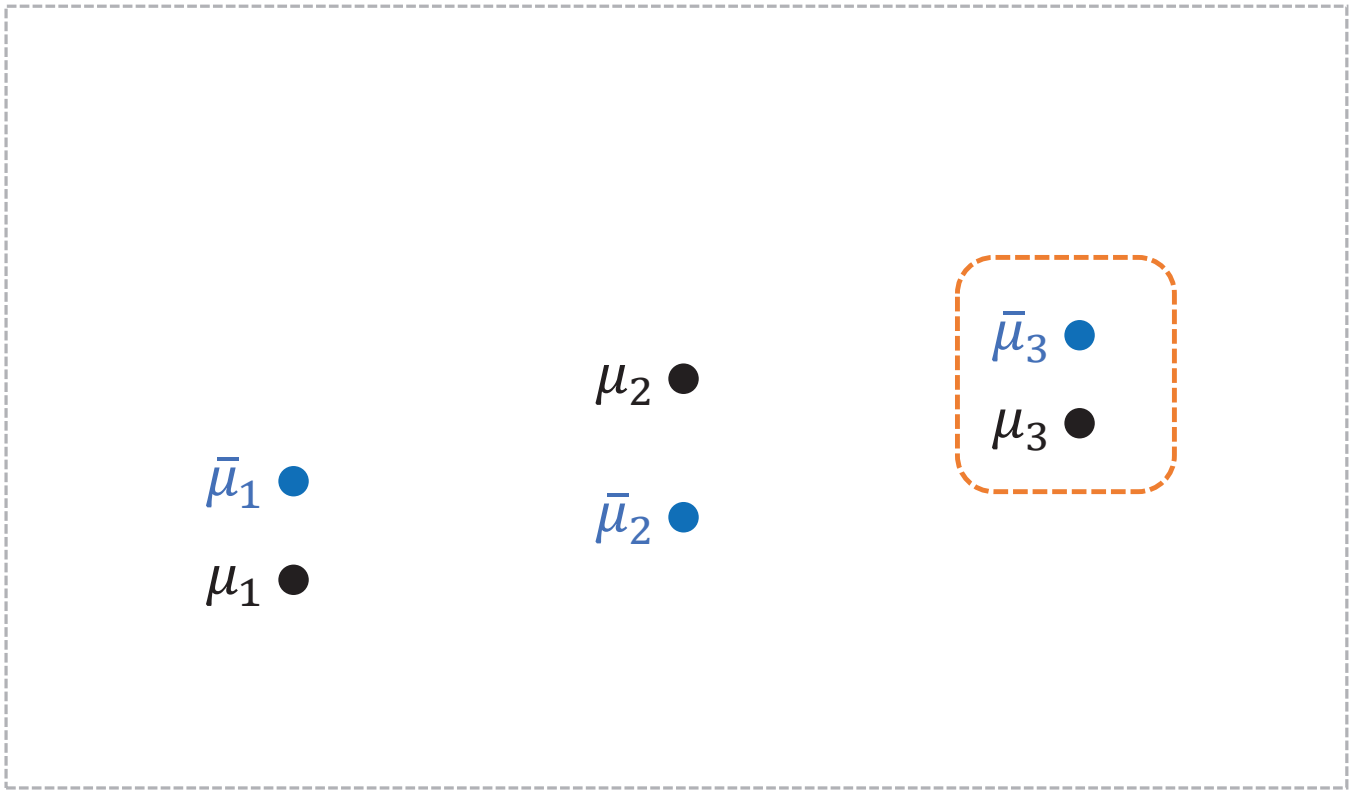
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



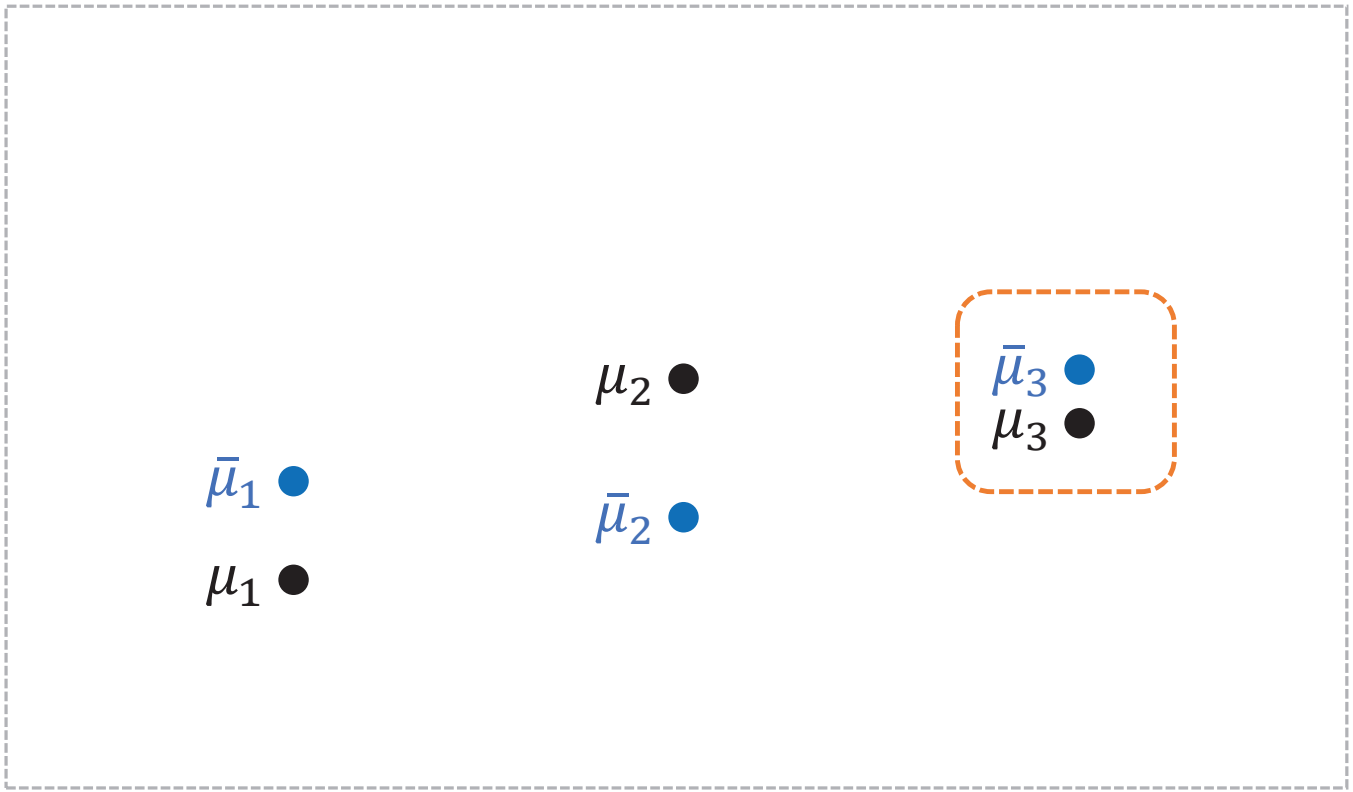
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



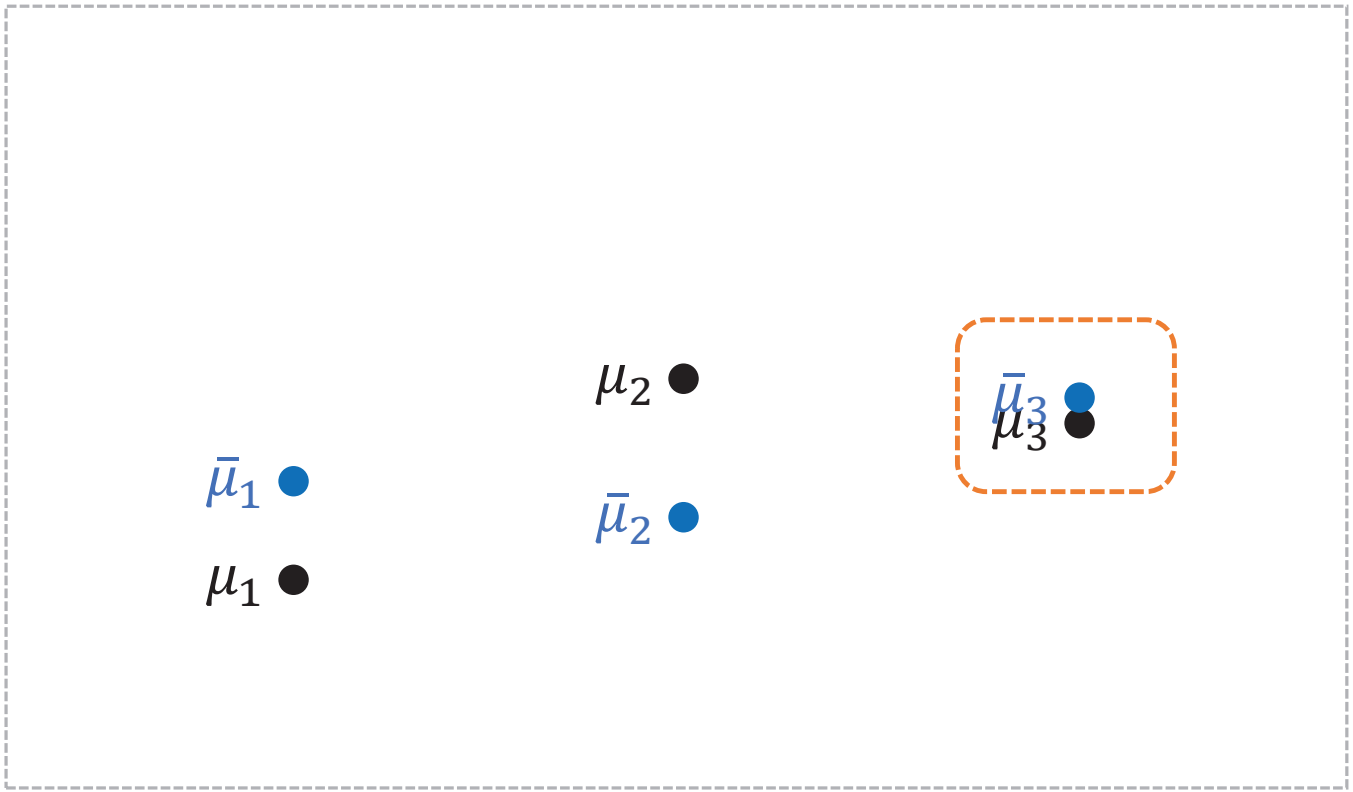
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



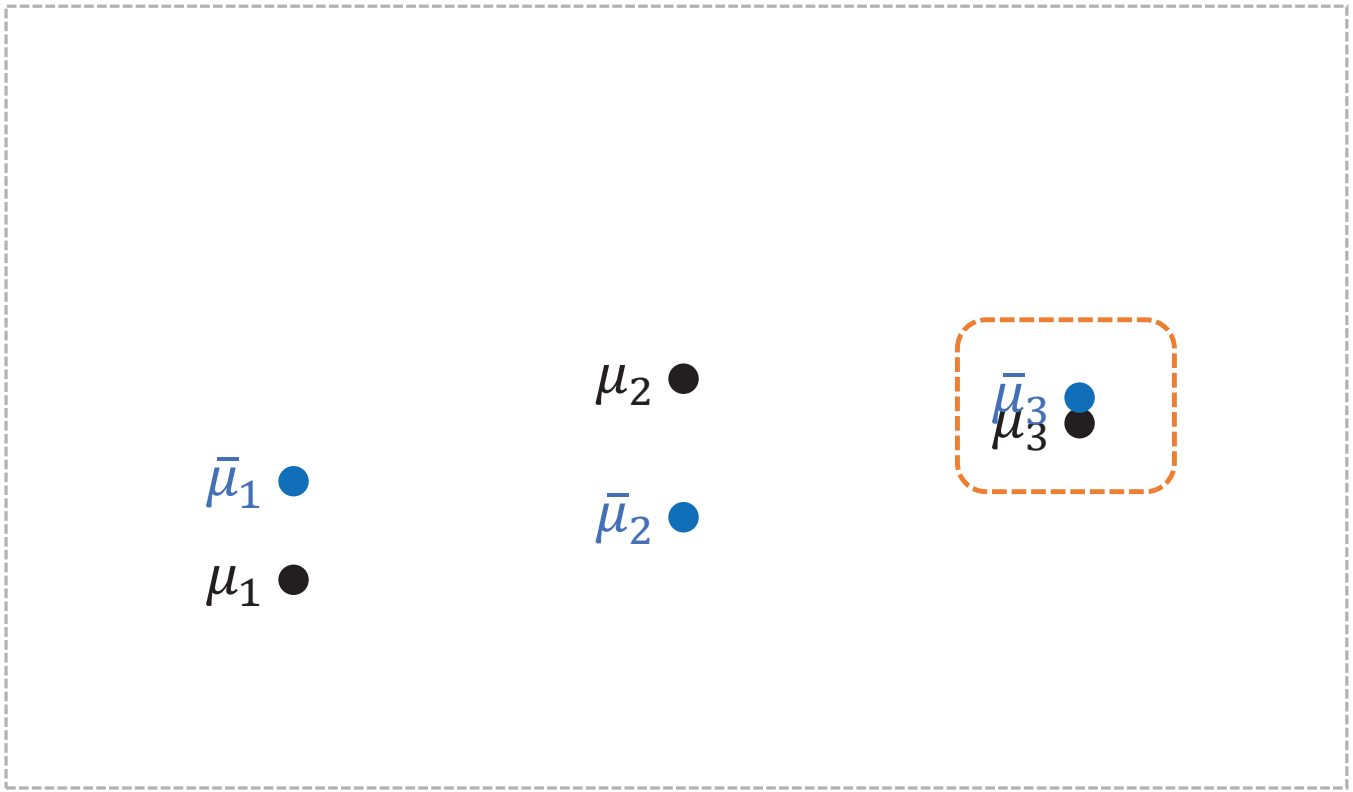
Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean



Upper Confidence Bound (UCB)

- A Naive Approach based on Sample Mean

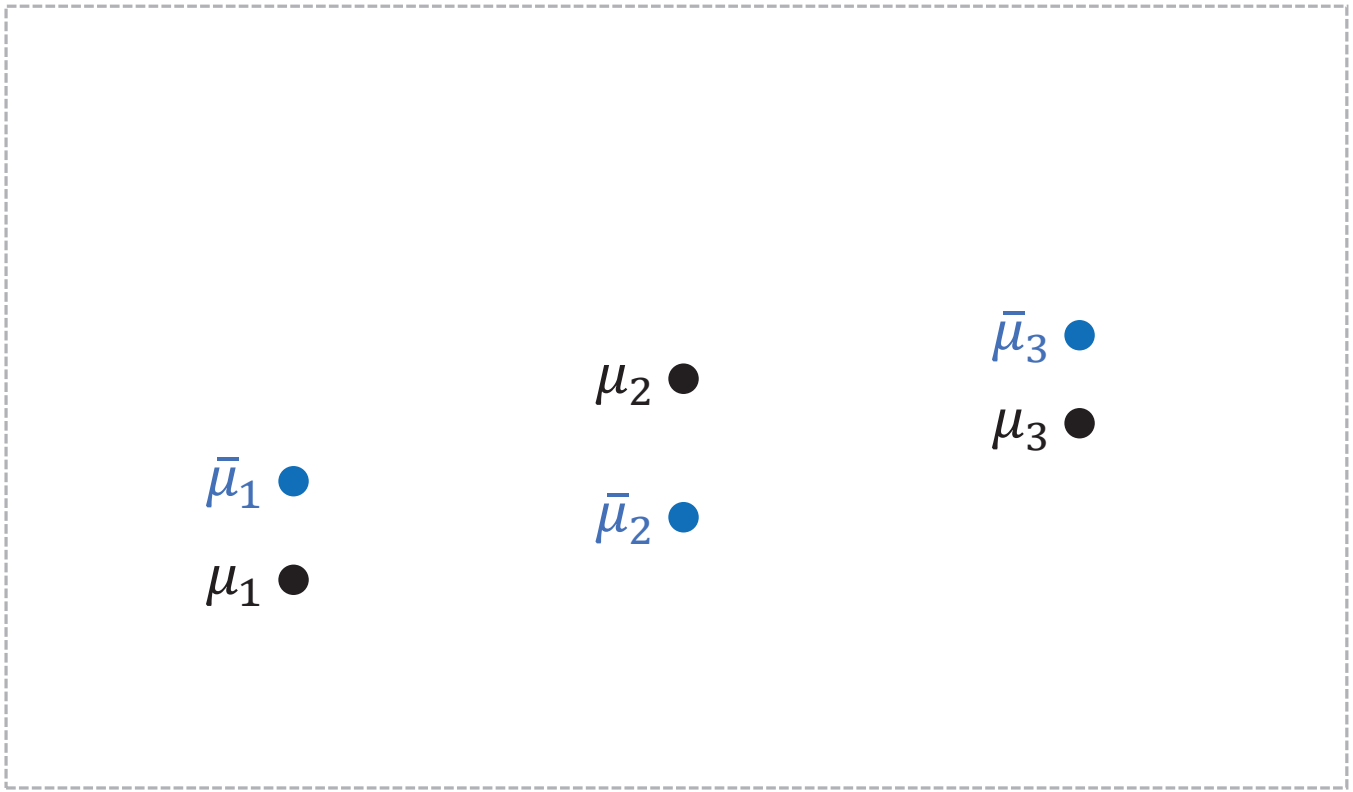


- A bad regret bound

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

Upper Confidence Bound (UCB)

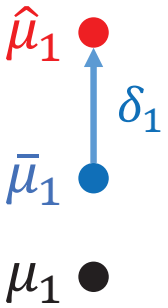
- The Algorithm of UCB

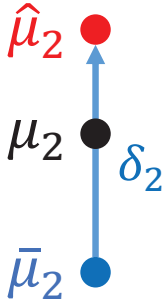


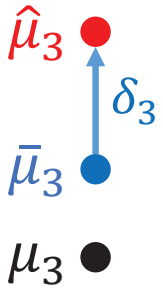
Upper Confidence Bound (UCB)

- The Algorithm of UCB

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



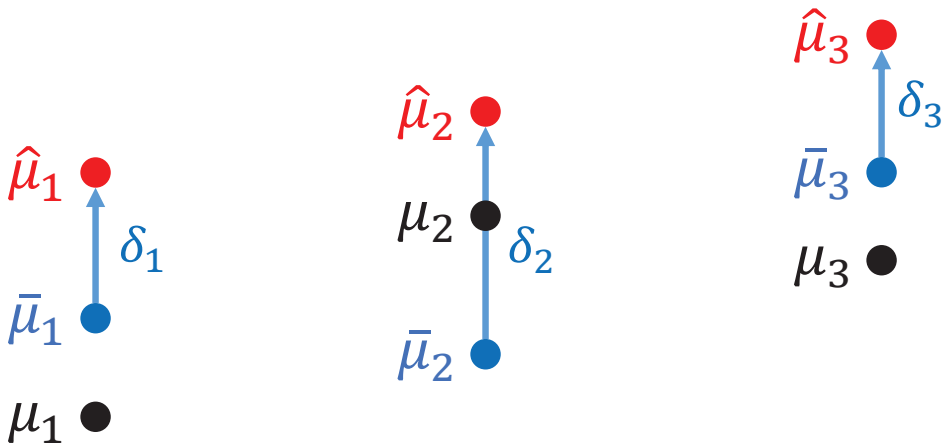




Upper Confidence Bound (UCB)

- The Algorithm of UCB

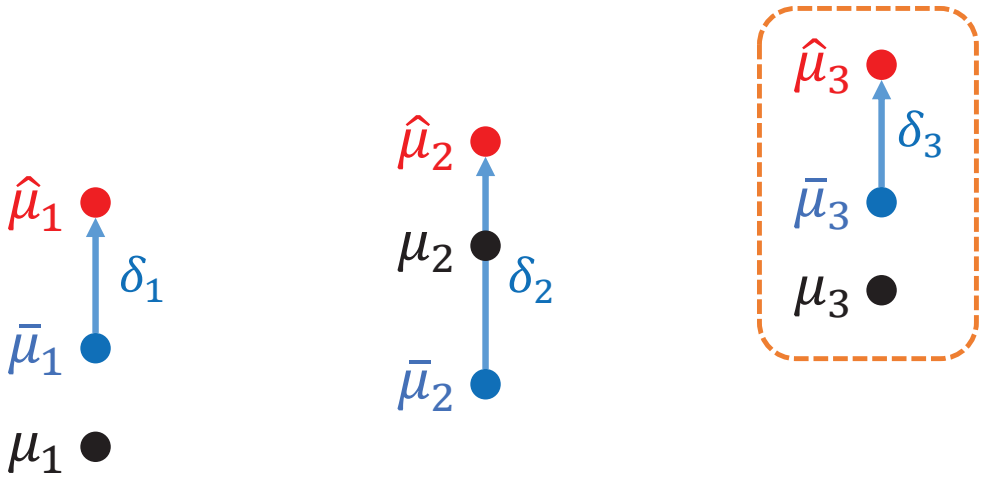
With high probability $\mu_i \leq \hat{\mu}_i = \bar{\mu}_i + \delta_i$



Upper Confidence Bound (UCB)

- The Algorithm of UCB

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

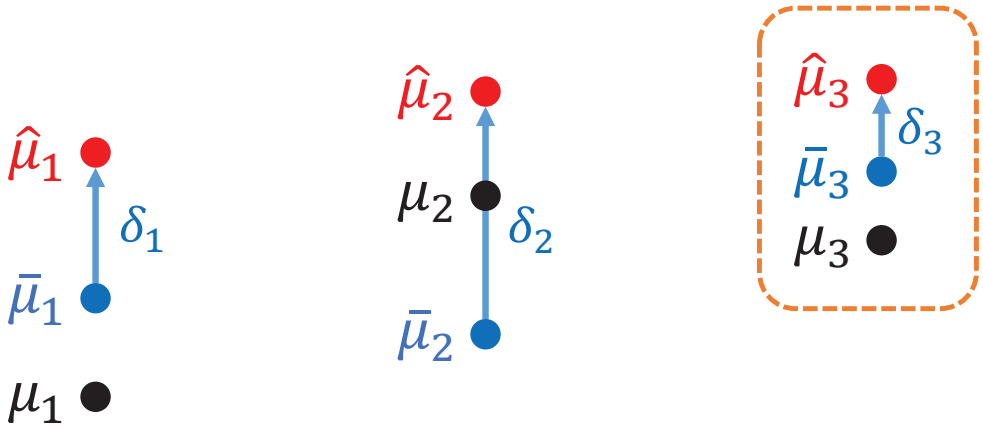


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \hat{\mu}_i$

Upper Confidence Bound (UCB)

- The Algorithm of UCB

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

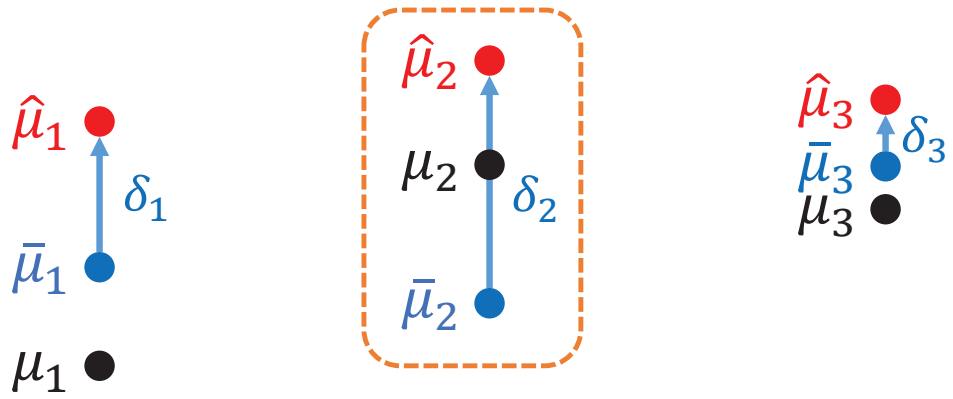


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \hat{\mu}_i$

Upper Confidence Bound (UCB)

The Algorithm of UCB

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

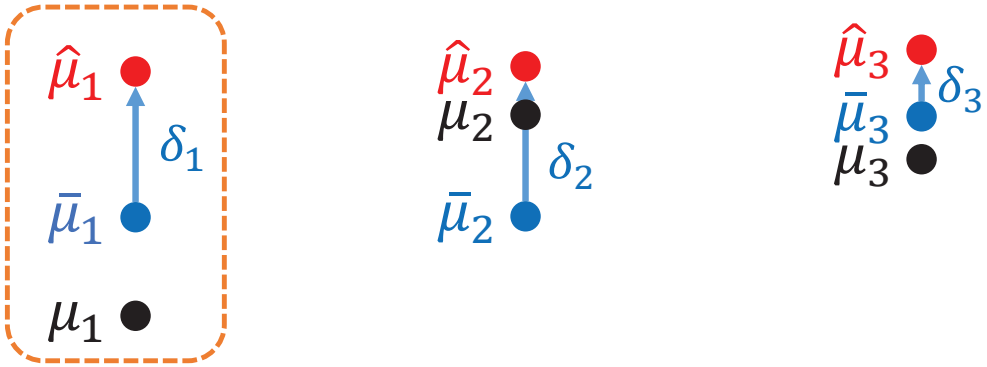


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \hat{\mu}_i$

Upper Confidence Bound (UCB)

- The Algorithm of UCB

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

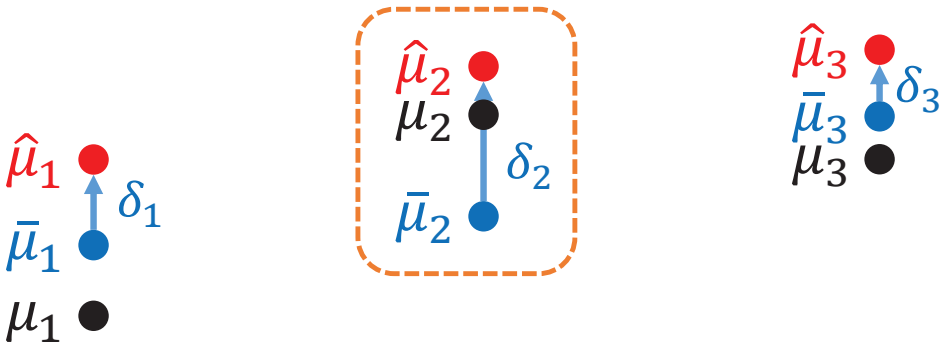


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \bar{\mu}_i$

Upper Confidence Bound (UCB)

The Algorithm of UCB

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

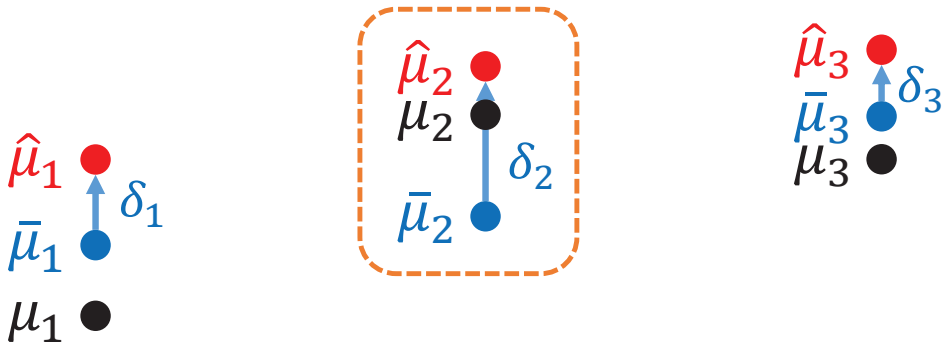


Optimism in Face of Uncertainty: $i_t = \text{argmax}_i \, \bar{\mu}_i$

Upper Confidence Bound (UCB)

The Algorithm of UCB

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



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

- Construct $\bar{\mu}_i$ by concentration inequalities [Auer et al., 2002a]

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

Hoeffding-based UCB

- The Algorithm

1: Play each machine once
2: **for** $t = K + 1, 2, \ldots, T$ **do**

6: **end for**

Hoeffding-based UCB

- The Algorithm

1: Play each machine once
2: **for** $t = K + 1, 2, \ldots, T$ **do**
3: Play

$$i_t = \underset{i \in [K]}{\operatorname{argmax}} \bar{\mu}_i(n_i^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_i^{t-1}}}$$

and observe $X_{i_t}^t$

6: **end for**

Hoeffding-based UCB

- The Algorithm

1: Play each machine once

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

3: Play

$$i_t = \underset{i \in [K]}{\operatorname{argmax}} \bar{\mu}_i(n_i^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_i^{t-1}}}$$

and observe $X_{i_t}^t$

4:

$$n_{i_t}^t = n_{i_t}^{t-1} + 1, \quad n_i^t = n_i^{t-1},\ i \neq i_t$$

5:

$$\bar{\mu}_{i_t}(n_{i_t}^t) = \frac{n_{i_t}^{t-1} \times \bar{\mu}_{i_t}(n_{i_t}^{t-1}) + X_{i_t}^t}{n_{i_t}^t}$$

6: **end for**

Analysis I

Theorem 1 (Chernoff-Hoeffding bound [Auer et al., 2002a])

Let $X_1, \ldots, X_T$ be random variables with common range $[0, 1]$ and such that

$$\mathrm{E}\left[X_t | X_1, \ldots, X_{t-1}\right] = \mu$$

Let $\bar{X} = \sum_{t=1}^{T} X_t / T$. Then for all $\alpha \geq 0$,

$$\Pr\left[\bar{X} \geq \mu + \alpha\right] \leq e^{-2T\alpha^2}, \ \Pr\left[\bar{X} \leq \mu - \alpha\right] \leq e^{-2T\alpha^2}$$

Analysis I

Theorem 1 (Chernoff-Hoeffding bound [Auer et al., 2002a])

Let $X_1, \ldots, X_T$ be random variables with common range $[0, 1]$ and such that

$$\mathrm{E}\left[X_t | X_1, \ldots, X_{t-1}\right] = \mu$$

Let $\bar{X} = \sum_{t=1}^{T} X_t / T$. Then for all $\alpha \geq 0$,

$$\Pr\left[\bar{X} \geq \mu + \alpha\right] \leq e^{-2T\alpha^2}, \ \Pr\left[\bar{X} \leq \mu - \alpha\right] \leq e^{-2T\alpha^2}$$

- Notations

$$* = \operatorname*{argmax}_{i \in [K]} \mu_i, \quad \mu_* = \max_{i \in [K]} \mu_i, \quad \Delta_i = \mu_* - \mu_i$$

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

LAMDA Learning And Mining from DatA

Analysis II

$$n_i^T = 1 + \sum_{t=K+1}^{T} \{i_t = i\} \leq \ell + \sum_{t=K+1}^{T} \left\{ i_t = i, n_i^{t-1} \geq \ell \right\}$$

Analysis II

$$n_i^T = 1 + \sum_{t=K+1}^{T} \{i_t = i\} \leq \ell + \sum_{t=K+1}^{T} \left\{ i_t = i, n_i^{t-1} \geq \ell \right\}$$

$$\leq \ell + \sum_{t=K+1}^{T} \left\{ \bar{\mu}_*(n_*^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_*^{t-1}}} \leq \bar{\mu}_i(n_i^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_i^{t-1}}}, n_i^{t-1} \geq \ell \right.$$

Analysis II

$$n_i^T = 1 + \sum_{t=K+1}^{T} \{i_t = i\} \leq \ell + \sum_{t=K+1}^{T} \left\{ i_t = i, n_i^{t-1} \geq \ell \right\}$$

$$\leq \ell + \sum_{t=K+1}^{T} \left\{ \bar{\mu}_*(n_*^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_*^{t-1}}} \leq \bar{\mu}_i(n_i^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_i^{t-1}}}, n_i^{t-1} \geq \ell \right.$$

$$\leq \ell + \sum_{t=K+1}^{T} \left\{ \min_{0<p<t} \bar{\mu}_*(p) + \sqrt{\frac{2\log(t-1)}{p}} \leq \max_{\ell \leq q < t} \bar{\mu}_i(q) + \sqrt{\frac{2\log(t-1)}{q}} \right\}$$

Analysis II

$$n_i^T = 1 + \sum_{t=K+1}^{T} \{i_t = i\} \leq \ell + \sum_{t=K+1}^{T} \left\{ i_t = i, n_i^{t-1} \geq \ell \right\}$$

$$\leq \ell + \sum_{t=K+1}^{T} \left\{ \bar{\mu}_*(n_*^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_*^{t-1}}} \leq \bar{\mu}_i(n_i^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_i^{t-1}}}, n_i^{t-1} \geq \ell \right.$$

$$\leq \ell + \sum_{t=K+1}^{T} \left\{ \min_{0<p<t} \bar{\mu}_*(p) + \sqrt{\frac{2\log(t-1)}{p}} \leq \max_{\ell \leq q < t} \bar{\mu}_i(q) + \sqrt{\frac{2\log(t-1)}{q}} \right\}$$

$$\leq \ell + \sum_{t=1}^{T-1} \sum_{p=1}^{t-1} \sum_{q=\ell}^{t-1} \left\{ \bar{\mu}_*(p) + \sqrt{\frac{2\log(t)}{p}} \leq \bar{\mu}_i(q) + \sqrt{\frac{2\log(t)}{q}} \right\}$$

Analysis II

$$n_i^T = 1 + \sum_{t=K+1}^{T} \{i_t = i\} \leq \ell + \sum_{t=K+1}^{T} \left\{ i_t = i, n_i^{t-1} \geq \ell \right\}$$

$$\leq \ell + \sum_{t=K+1}^{T} \left\{ \bar{\mu}_*(n_*^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_*^{t-1}}} \leq \bar{\mu}_i(n_i^{t-1}) + \sqrt{\frac{2\log(t-1)}{n_i^{t-1}}}, n_i^{t-1} \geq \ell \right.$$

$$\leq \ell + \sum_{t=K+1}^{T} \left\{ \min_{0<p<t} \bar{\mu}_*(p) + \sqrt{\frac{2\log(t-1)}{p}} \leq \max_{\ell \leq q < t} \bar{\mu}_i(q) + \sqrt{\frac{2\log(t-1)}{q}} \right\}$$

$$\leq \ell + \sum_{t=1}^{T-1} \sum_{p=1}^{t-1} \sum_{q=\ell}^{t-1} \left\{ \bar{\mu}_*(p) + \sqrt{\frac{2\log(t)}{p}} \leq \bar{\mu}_i(q) + \sqrt{\frac{2\log(t)}{q}} \right\}$$

$$\leq \ell + \sum_{t=1}^{T-1} \sum_{p=1}^{t-1} \sum_{q=\ell}^{t-1} \left(\left\{ \bar{\mu}_*(p) + \sqrt{\frac{2\log(t)}{p}} \leq \mu_* \right\} \right.$$

$$\left. + \left\{ \mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}} \right\} + \left\{ \mu_i + \sqrt{\frac{2\log(t)}{q}} \leq \bar{\mu}_i(q) \right\} \right)$$

Analysis III

$$\mathrm{E}[n_i^T] \leq \ell + \sum_{t=1}^{T-1} \sum_{p=1}^{t-1} \sum_{q=\ell}^{t-1} \left(\Pr\left[\bar{\mu}_*(p) + \sqrt{\frac{2\log(t)}{p}} \leq \mu_* \right] \right.$$
$$\left. + \Pr\left[\mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}} \right] + \Pr\left[\mu_i + \sqrt{\frac{2\log(t)}{q}} \leq \bar{\mu}_i(q) \right] \right)$$

Analysis III

$$\begin{aligned}
\mathrm{E}[n_i^T] \leq & \ell + \sum_{t=1}^{T-1} \sum_{p=1}^{t-1} \sum_{q=\ell}^{t-1} \left(\Pr\left[\bar{\mu}_*(p) + \sqrt{\frac{2\log(t)}{p}} \leq \mu_* \right] \right. \\
& \left. + \Pr\left[\mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}} \right] + \Pr\left[\mu_i + \sqrt{\frac{2\log(t)}{q}} \leq \bar{\mu}_i(q) \right] \right) \\
\leq & \ell + \sum_{t=1}^{T-1} \sum_{p=1}^{t-1} \sum_{q=\ell}^{t-1} \left(t^{-4} + \Pr\left[\mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}} \right] + t^{-4} \right)
\end{aligned}$$

Analysis III

$$
\begin{aligned}
\mathrm{E}[n_i^T] \leq & \ell + \sum_{t=1}^{T-1}\sum_{p=1}^{t-1}\sum_{q=\ell}^{t-1}\left(\Pr\left[\bar{\mu}_*(p) + \sqrt{\frac{2\log(t)}{p}} \leq \mu_*\right]\right. \\
& \left. + \Pr\left[\mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}}\right] + \Pr\left[\mu_i + \sqrt{\frac{2\log(t)}{q}} \leq \bar{\mu}_i(q)\right]\right) \\
\leq & \ell + \sum_{t=1}^{T-1}\sum_{p=1}^{t-1}\sum_{q=\ell}^{t-1}\left(t^{-4} + \Pr\left[\mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}}\right] + t^{-4}\right) \\
= & \left\lceil \frac{8\ln T}{\Delta_i^2} \right\rceil + 2\sum_{t=1}^{\infty}\sum_{p=1}^{t-1}\sum_{q=\lceil 8\ln T/\Delta_i^2\rceil}^{t-1} t^{-4}
\end{aligned}
$$

Analysis III

$$
\begin{aligned}
\mathrm{E}[n_i^T] \leq & \ell + \sum_{t=1}^{T-1}\sum_{p=1}^{t-1}\sum_{q=\ell}^{t-1}\left(\Pr\left[\bar{\mu}_*(p) + \sqrt{\frac{2\log(t)}{p}} \leq \mu_*\right]\right. \\
& \left. + \Pr\left[\mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}}\right] + \Pr\left[\mu_i + \sqrt{\frac{2\log(t)}{q}} \leq \bar{\mu}_i(q)\right]\right) \\
\leq & \ell + \sum_{t=1}^{T-1}\sum_{p=1}^{t-1}\sum_{q=\ell}^{t-1}\left(t^{-4} + \Pr\left[\mu_* \leq \mu_i + 2\sqrt{\frac{2\log(t)}{q}}\right] + t^{-4}\right) \\
= & \left\lceil \frac{8\ln T}{\Delta_i^2} \right\rceil + 2\sum_{t=1}^{\infty}\sum_{p=1}^{t-1}\sum_{q=\lceil 8\ln T/\Delta_i^2\rceil}^{t-1} t^{-4} \\
\leq & \frac{8\ln T}{\Delta_i^2} + 1 + 2\sum_{t=1}^{\infty} t^{-2} \leq \frac{8\ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}
\end{aligned}
$$

LAMDA Learning And Mining from DatA

Analysis IV

$$\begin{aligned}
\mathrm{E}\left[\mathrm{Regret}\right] &= \sum_{i \neq *} \Delta_i \mathrm{E}\left[n_i^T\right] \\
&\leq 8 \sum_{i \neq *} \frac{\ln T}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \sum_{i \neq *} \Delta_i \\
&= O\left(K \log T\right)
\end{aligned}$$

Analysis IV

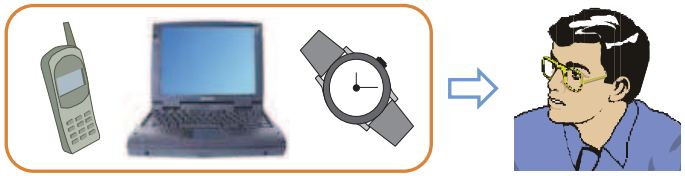
$$\begin{aligned}
\mathrm{E}\left[\mathrm{Regret}\right] &= \sum_{i \neq *} \Delta_i \mathrm{E}\left[n_i^T\right] \\
&\leq 8 \sum_{i \neq *} \frac{\ln T}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \sum_{i \neq *} \Delta_i \\
&= O\left(K \log T\right)
\end{aligned}$$

- Extensions
  - High-probability Bound [Abbasi-yadkori et al., 2011]
  - Distribution-free Bound [Bubeck and Cesa-Bianchi, 2012]
  - Non-stochastic MAB [Auer et al., 2002b]

Outline

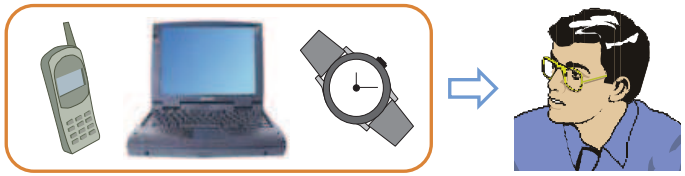
Stochastic Linear Bandits (SLB)

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



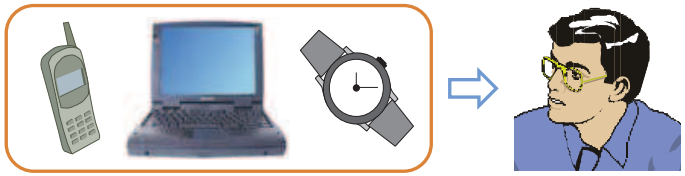
Stochastic Linear Bandits (SLB)

- 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



Stochastic Linear Bandits (SLB)

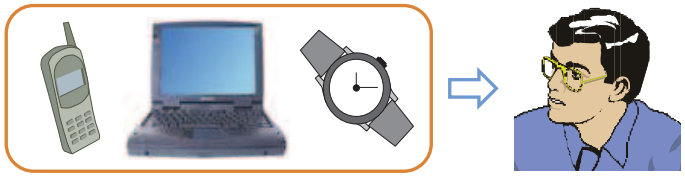
- Recommendation by Multi-Armed Bandits (MAB)
  - Each item is an arm
  - User feedback is the reward
  - The $O(K \log T)$ regret bound is loose for large K



- Stochastic Linear Bandits (SLB)
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$

Stochastic Linear Bandits (SLB)

- 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



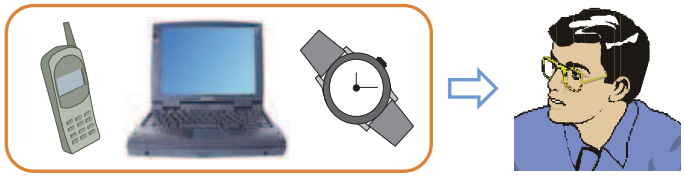
- Stochastic Linear Bandits (SLB)
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, reward $\mu_{\mathbf{x}}$ and feedback y satisfy

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* = \mathrm{E}[y|\mathbf{x}] \Leftrightarrow y = \mathbf{x}^\top \mathbf{w}_* + \epsilon$$

    where ϵ is a zero-mean random noise

Stochastic Linear Bandits (SLB)

- 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



- Stochastic Linear Bandits (SLB)
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, reward $\mu_{\mathbf{x}}$ and feedback y satisfy

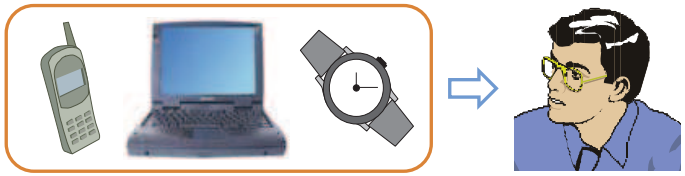
$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* = \mathrm{E}[y|\mathbf{x}] \Leftrightarrow y = \mathbf{x}^\top \mathbf{w}_* + \epsilon$$

    where ϵ is a zero-mean random noise
  - Select a sequence of arms $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathcal{D}$ to minimize

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

Stochastic Linear Bandits (SLB)

- 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



- Stochastic Linear Bandits (SLB)
  - Each arm is a feature vector $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$
  - For arm $\mathbf{x}$, reward $\mu_{\mathbf{x}}$ and feedback y satisfy

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* = \mathrm{E}[y|\mathbf{x}] \Leftrightarrow y = \mathbf{x}^\top \mathbf{w}_* + \epsilon$$

    where ϵ is a zero-mean random noise
  - Select a sequence of arms $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathcal{D}$ to minimize

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

  - The UCB algorithm also be applied [Dani et al., 2008b]

The Algorithm

In the t-th Round

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

The Algorithm

In the t-th Round

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

Learning History

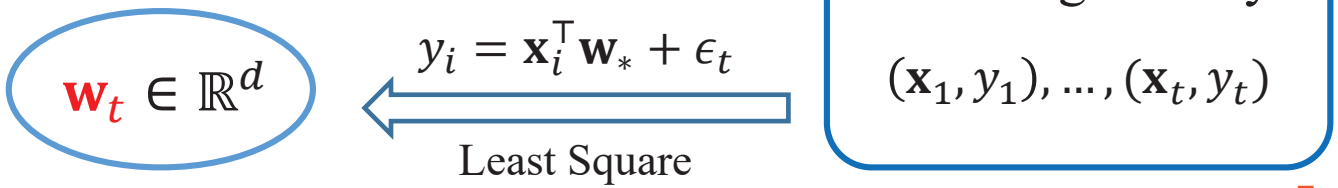
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)$

LAMDA Learning And Mining from DatA

The Algorithm

In the t-th Round

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

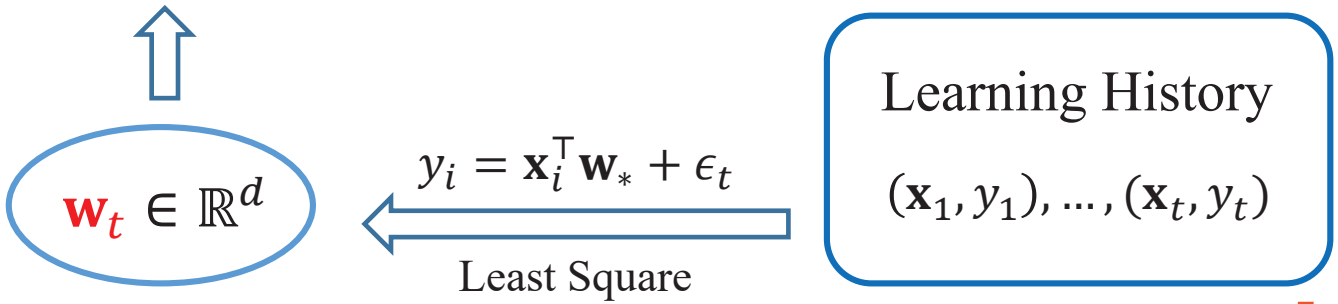


The Algorithm

In the t-th Round

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

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



LAMDA
Learning And Mining from DatA

The Algorithm

In the t-th Round

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

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

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

$$\mathbf{w}_t \in \mathbb{R}^d \quad \xleftarrow[\text{Least Square}]{y_i = \mathbf{x}_i^\top \mathbf{w}_* + \epsilon_t} \quad \text{Learning History } (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)$$

LAMDA Learning And Mining from DatA

The Algorithm

In the t-th Round

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

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* \leq \underbrace{\max_{\mathbf{w} \in \mathcal{C}_t} \mathbf{x}^\top \mathbf{w} = \hat{\mu}_{\mathbf{x}}}_{\text{Upper Bound}} \Longrightarrow \mathbf{x}_{t+1} = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \hat{\mu}_{\mathbf{x}}$$

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

$$\mathbf{w}_t \in \mathbb{R}^d \quad \Longleftarrow \underbrace{y_i = \mathbf{x}_i^\top \mathbf{w}_* + \epsilon_t}_{\text{Least Square}} \quad \text{Learning History } (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)$$

LAMDA Learning And Mining from DatA

The Algorithm

In the t-th Round

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

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* \leq \underbrace{\max_{\mathbf{w} \in \mathcal{C}_t} \mathbf{x}^\top \mathbf{w} = \hat{\mu}_{\mathbf{x}}}_{\text{Upper Bound}} \Longrightarrow \mathbf{x}_{t+1} = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \hat{\mu}_{\mathbf{x}}$$

Submit $\mathbf{x}_{t+1}$

Observe $y_{t+1} \in \mathbb{R}$

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

Learning History

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)$

$y_i = \mathbf{x}_i^\top \mathbf{w}_* + \epsilon_t$

Least Square

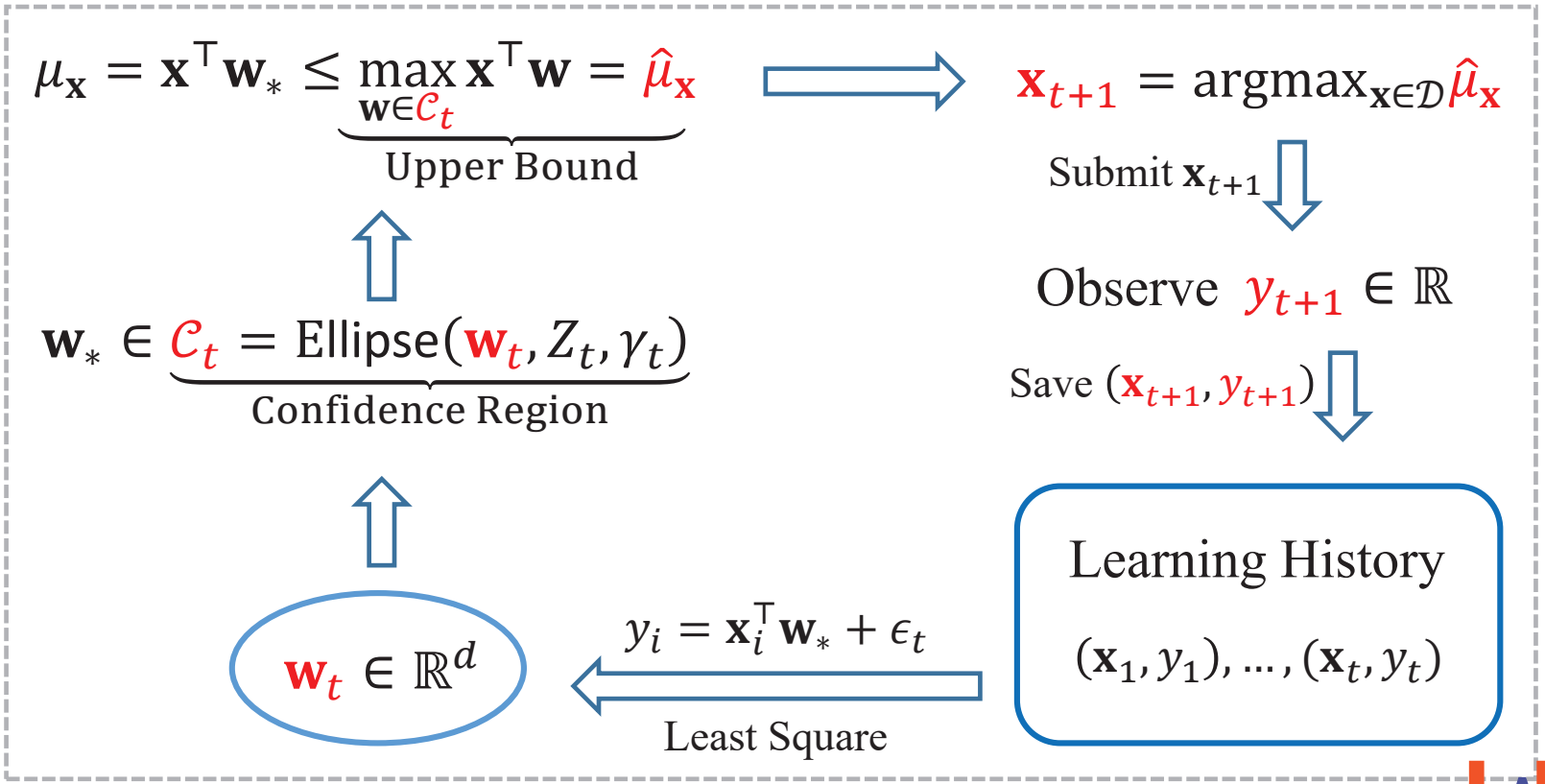
$\mathbf{w}_t \in \mathbb{R}^d$

LAMDA Learning And Mining from DatA

The Algorithm

In the t-th Round

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



LAMDA
Learning And Mining from DatA

The Algorithm

In the t-th Round

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

$$\mu_{\mathbf{x}} = \mathbf{x}^\top \mathbf{w}_* \leq \underbrace{\max_{\mathbf{w} \in \mathcal{C}_t} \mathbf{x}^\top \mathbf{w} = \hat{\mu}_{\mathbf{x}}}_{\text{Upper Bound}} \Longrightarrow \mathbf{x}_{t+1} = \text{argmax}_{\mathbf{x} \in \mathcal{D}} \hat{\mu}_{\mathbf{x}}$$

Submit $\mathbf{x}_{t+1}$

Observe $y_{t+1} \in \mathbb{R}$

$(\mathbf{x}_{t+1}, y_{t+1})$

Partial History

$\hat{\mathbf{w}}_{t-1}, Z_{t-1}, (\mathbf{x}_t, y_t)$

Online Least Square $y_i = \mathbf{x}_i^\top \mathbf{w}_* + \epsilon_t$

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

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

LAMDA Learning And Mining from DatA

Implementation Issues

- Least Square [Abbasi-yadkori et al., 2011]

$$\mathbf{w}_t = \left(\lambda I + \sum_{i=1}^{t} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i=1}^{t} y_i \mathbf{x}_i\right)$$

Implementation Issues

- Least Square [Abbasi-yadkori et al., 2011]

$$\mathbf{w}_t = \left(\lambda I + \sum_{i=1}^{t} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i=1}^{t} y_i \mathbf{x}_i\right)$$

- Online Least Square

$$\mathbf{w}_t = \left(Z_{t-1}^{-1} - \frac{Z_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top Z_{t-1}^{-1}}{1+\mathbf{x}_t^\top Z_{t-1}^{-1}\mathbf{x}_t}\right)(\mathbf{z}_{t-1} + y_t\mathbf{x}_t)$$

$$= \mathbf{w}_{t-1} + \left(y_t - \frac{\mathbf{x}_t^\top Z_{t-1}^{-1}\mathbf{z}_t}{1+\mathbf{x}_t^\top Z_{t-1}^{-1}\mathbf{x}_t}\right) Z_{t-1}^{-1}\mathbf{x}_t$$

where $Z_{t-1} = \lambda I + \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i^\top$ and $\mathbf{z}_{t-1} = \sum_{i=1}^{t-1} y_i \mathbf{x}_i$

Implementation Issues

- Least Square [Abbasi-yadkori et al., 2011]

$$\mathbf{w}_t = \left(\lambda I + \sum_{i=1}^{t} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i=1}^{t} y_i \mathbf{x}_i\right)$$

- Online Least Square

$$\mathbf{w}_t = \left(Z_{t-1}^{-1} - \frac{Z_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top Z_{t-1}^{-1}}{1+\mathbf{x}_t^\top Z_{t-1}^{-1}\mathbf{x}_t}\right)(\mathbf{z}_{t-1} + y_t\mathbf{x}_t)$$

$$= \mathbf{w}_{t-1} + \left(y_t - \frac{\mathbf{x}_t^\top Z_{t-1}^{-1}\mathbf{z}_t}{1+\mathbf{x}_t^\top Z_{t-1}^{-1}\mathbf{x}_t}\right) Z_{t-1}^{-1}\mathbf{x}_t$$

where $Z_{t-1} = \lambda I + \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i^\top$ and $\mathbf{z}_{t-1} = \sum_{i=1}^{t-1} y_i \mathbf{x}_i$

- Arm Selection [Dani et al., 2008b]

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

  - Tractable when $\mathcal{D}$ is discrete or a ball

Regret Bound [Abbasi-yadkori et al., 2011]

Theorem 2 (Confidence Region)

With a high probability, we have

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

for all $t > 0$.

- Self-Normalized Tail Inequality for Vector-Valued Martingales

Regret Bound [Abbasi-yadkori et al., 2011]

Theorem 2 (Confidence Region)

With a high probability, we have

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

for all $t > 0$.

- Self-Normalized Tail Inequality for Vector-Valued Martingales

Theorem 3 (Regret)

With a high probability, we have

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

for all $T > 0$.

LAMDA Learning And Mining from DatA

Extensions

- Non-stochastic Linear Bandits
  - [Dani et al., 2008a]
  - [Abernethy et al., 2008]
  - [Bartlett et al., 2008]
- Convex Bandits
  - [Flaxman et al., 2005]
  - [Saha and Tewari, 2011]
  - [Agarwal et al., 2010]

Reference I

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011).
Improved algorithms for linear stochastic bandits.
In *Advances in Neural Information Processing Systems 24*, pages 2312–2320.

Abernethy, J., Hazan, E., and Rakhlin, A. (2008).
Competing in the dark: An efficient algorithm for bandit linear optimization.
In *Proceedings of the 21st Annual Conference on Learning*, pages 263–274.

Agarwal, A., Dekel, O., and Xiao, L. (2010).
Optimal algorithms for online convex optimization with multi-point bandit feedback.
In *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 28–40.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a).
Finite-time analysis of the multiarmed bandit problem.
Machine Learning, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b).
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32(1):48–77.

LAMDA
Learning And Mining from DatA

Reference II


Bartlett, P. L., Dani, V., Hayes, T. P., Kakade, S. M., Rakhlin, A., and Tewari, A. (2008).
High-probability regret bounds for bandit online linear optimization.
In *Proceedings of the 21st Annual Conference on Learning*, pages 335–341.


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.


Dani, V., Hayes, T. P., and Kakade, S. M. (2008a).
The price of bandit information for online optimization.
In *Advances in Neural Information Processing Systems 20*, pages 345–352.


Dani, V., Hayes, T. P., and Kakade, S. M. (2008b).
Stochastic linear optimization under bandit feedback.
In *Proceedings of the 21st Annual Conference on Learning*, pages 355–366.


Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005).
Online convex optimization in the bandit setting: Gradient descent without a gradient.
In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394.

LAMDA
Learning And Mining from DatA

Reference III


Saha, A. and Tewari, A. (2011).
Improved regret guarantees for online smooth convex optimization with bandit feedback.
In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 636–642.


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