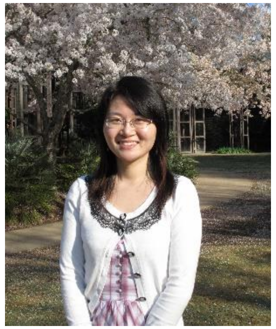
AN EFFICIENT PESSIMISTIC-OPTMISITIC ALGORITHM FOR CONSTRAINED STOCHASTIC LINEAR BANDITS

Xin Liu, Assistant Professor @ ShanghaiTech



Bin Li
Pennsylvania State University

Pengyi Shi
Purdue University

Lei Ying
University of Michigan, Ann Arbor

CONSTRAINED BANDITS: ONLINE DISPATCHING

Crowdsourcing

- ❑ Jobs arrive in a dynamic way
- ❑ Dispatch jobs to servers
- ❑ Observe reward, cost, and budget

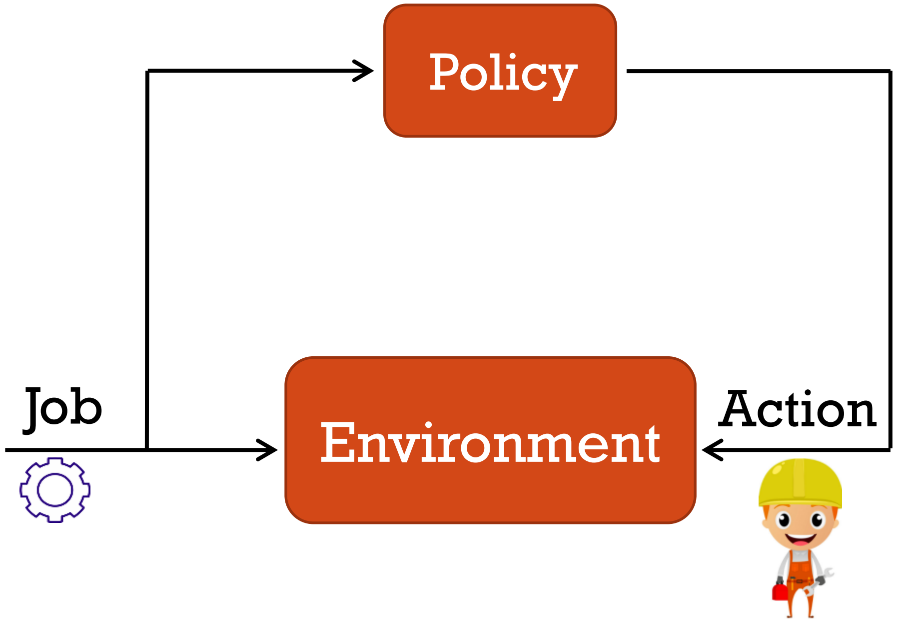
CONSTRAINED BANDITS: ONLINE DISPATCHING

Crowdsourcing

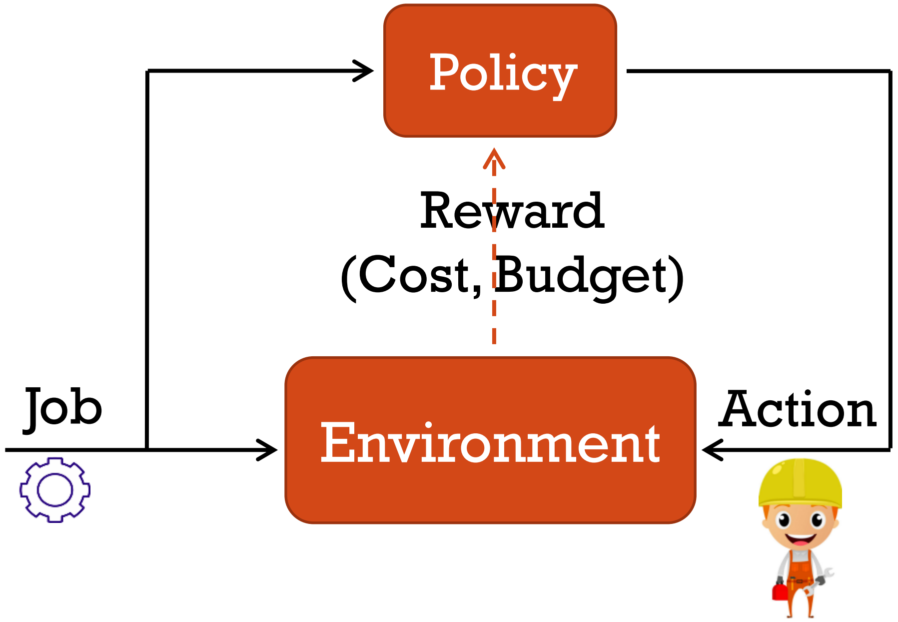
Healthcare

- ❑ Jobs arrive in a dynamic way
- ❑ Dispatch jobs to servers
- ❑ Observe reward, cost, and budget

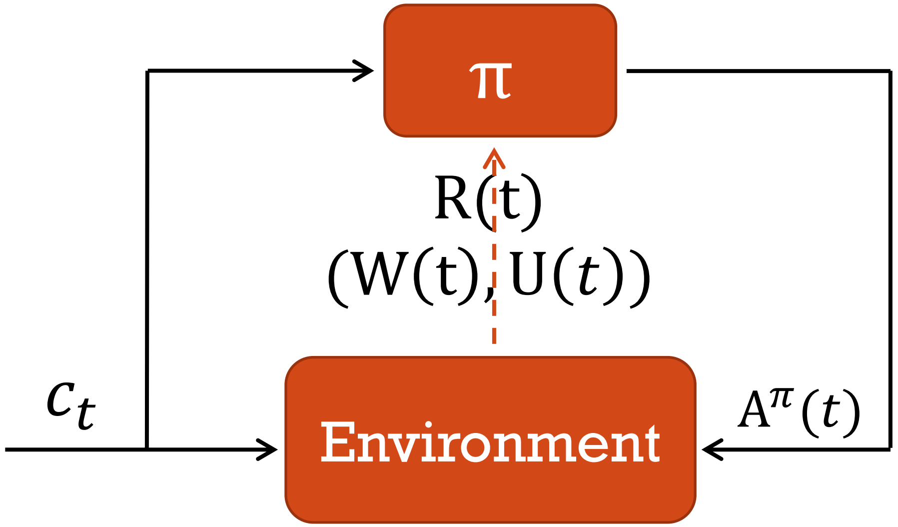
CONSTRAINED BANDITS: ONLINE DISPATCHING

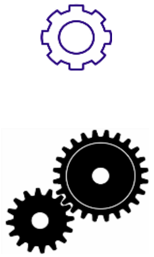


CONSTRAINED BANDITS: ONLINE DISPATCHING



CONSTRAINED BANDITS: ONLINE DISPATCHING



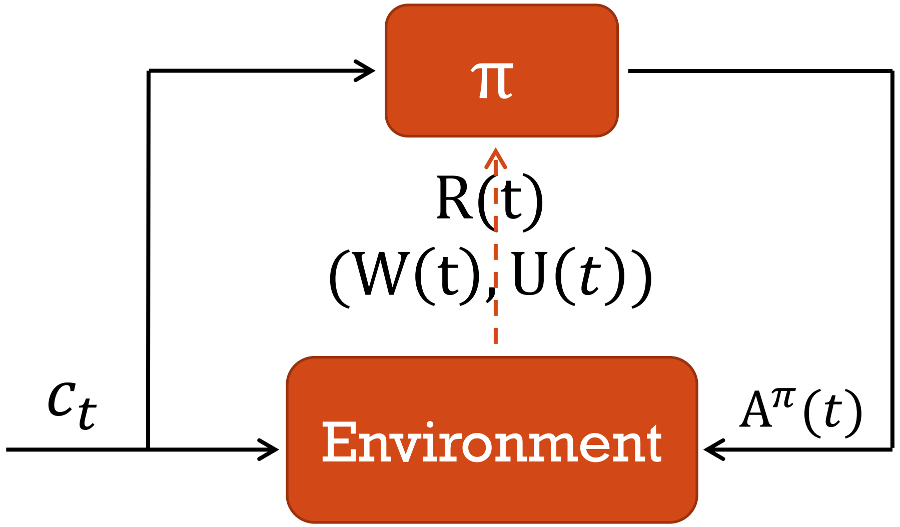


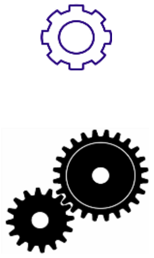
amazon mechanical turk

CONSTRAINED BANDITS: ONLINE DISPATCHING

$$\max_{\pi} \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{R}(t)\right]$$

$$\text{s.t.}\quad \mathrm{E}\left[\sum_{t=1}^{\tau} \mathrm{W}(t)\right] \le \mathrm{E}\left[\sum_{t=1}^{\tau} \mathrm{U}(t)\right]$$

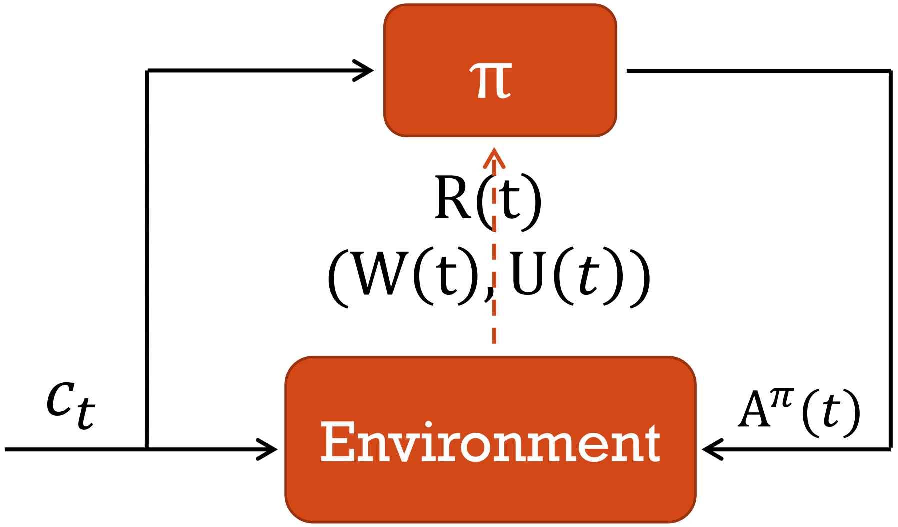




CONSTRAINED BANDITS: ONLINE DISPATCHING

$$\max_{\pi} \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{R}(c_t, \mathrm{A}^{\pi}(t))\right]$$

$$\text{s.t.} \quad \mathrm{E}\left[\sum_{t=1}^{\tau} \mathrm{W}(c_t, \mathrm{A}^{\pi}(t))\right] \leq \mathrm{E}\left[\sum_{t=1}^{\tau} \mathrm{U}(t)\right]$$



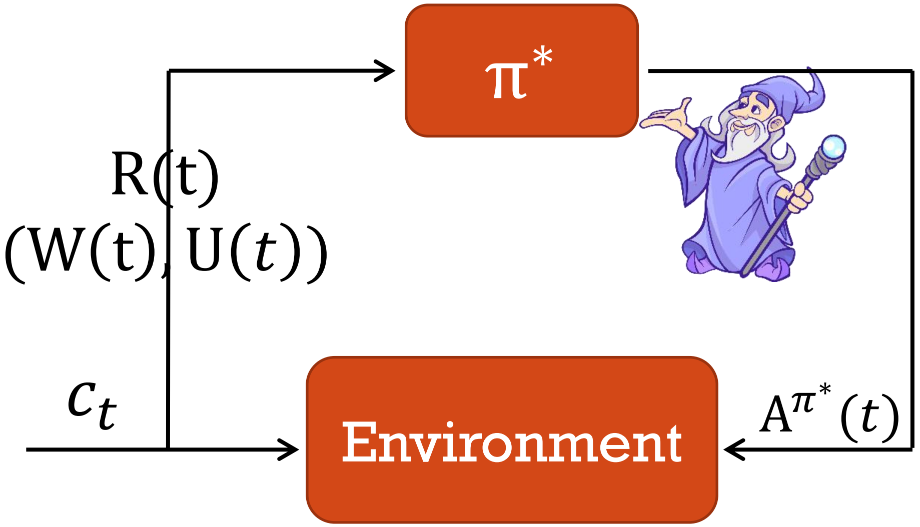
R( , ) = 0.1

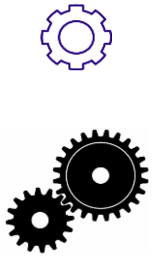
W( , ) = 90

CONSTRAINED BANDITS: ONLINE DISPATCHING

$$\max_{\pi} \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{R}(c_t, \mathrm{A}^{\pi}(t))\right]$$

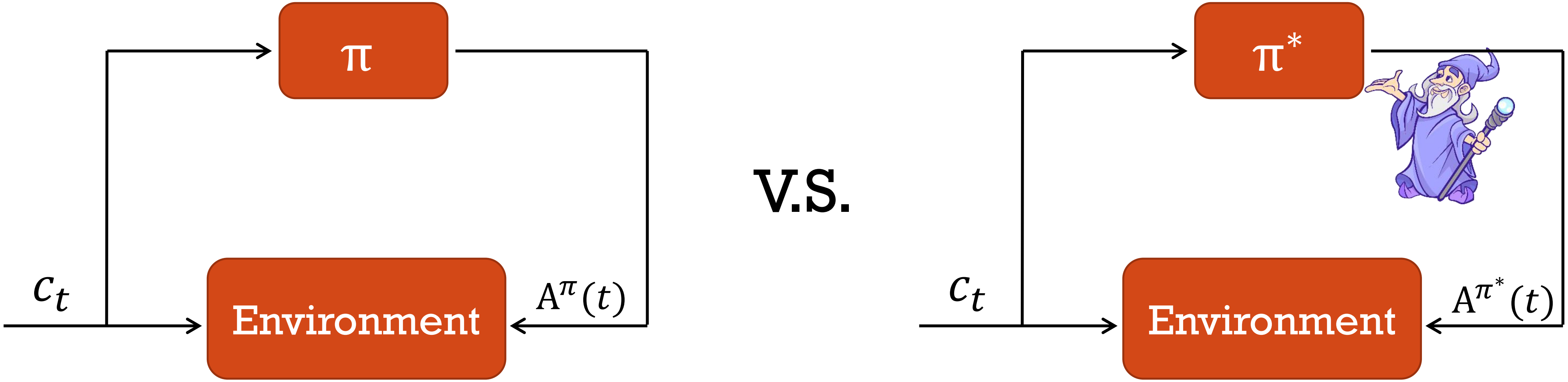
$$\text{s.t.} \quad \mathrm{E}\left[\sum_{t=1}^{\tau} \mathrm{W}(c_t, \mathrm{A}^{\pi}(t))\right] \leq \mathrm{E}\left[\sum_{t=1}^{\tau} \mathrm{U}(t)\right]$$





amazon mechanical turk

CONSTRAINED BANDITS: ONLINE DISPATCHING



Goals:

- ❑ Achieve optimal performance:

$$\text{Regret} = \text{TotalReward}(\pi^*) - \text{TotalReward}(\pi).$$

- ❑ Guarantee zero violation:

$$\text{Violation} = \text{TotalCost}(\pi) - \text{TotalBudget}.$$

CONSTRAINED STOCHASTIC LINEAR BANDITS

Model:

1. N servers and T time slots
2. A task arrives with feature c_t at time slot t.
3. Rewards are unknown:

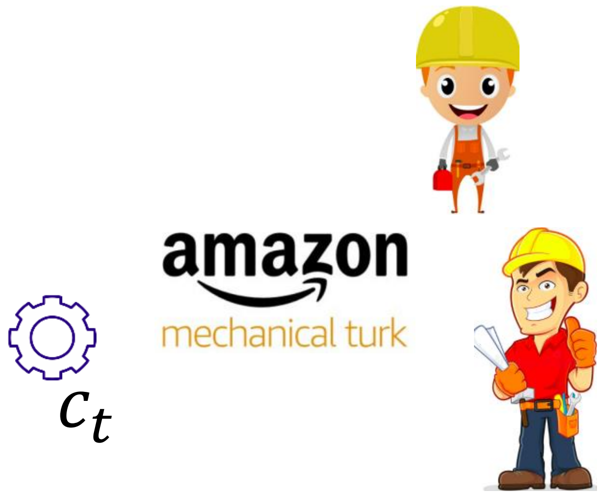
$$R(c_t, j) = < \theta_*, \phi(c_t, j) > + \eta_t$$

4. Costs are known:

$$W(c_t, j)$$

5. Stochastic (anytime) accumulative constraints:

$$E[\sum_{t=1}^{\tau} W(c_t, A(t))] \leq E[\sum_{t=1}^{\tau} U(t)]$$



CONSTRAINED STOCHASTIC LINEAR BANDITS

Model:

1. N servers and T time slots
2. A task arrives with feature c_t at time slot t.
3. Rewards are unknown:

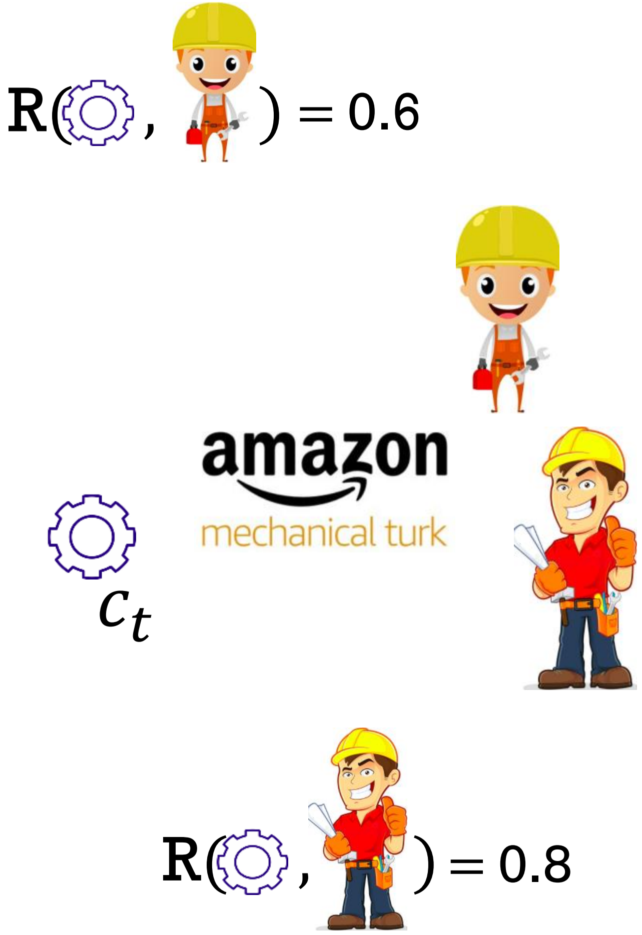
$$\mathrm{R}(c_t, j) = <\theta_*, \phi(c_t, j)> + \eta_t$$

4. Costs are known:

$$\mathrm{W}(c_t, j)$$

5. Stochastic (anytime) accumulative constraints:

$$\mathrm{E}[\sum_{t=1}^{\tau} \mathrm{W}(c_t, \mathrm{A}(t))] \leq \mathrm{E}[\sum_{t=1}^{\tau} \mathrm{U}(t)]$$



CONSTRAINED STOCHASTIC LINEAR BANDITS

Model:

1. N servers and T time slots
2. A task arrives with feature c_t at time slot t.
3. Rewards are unknown:

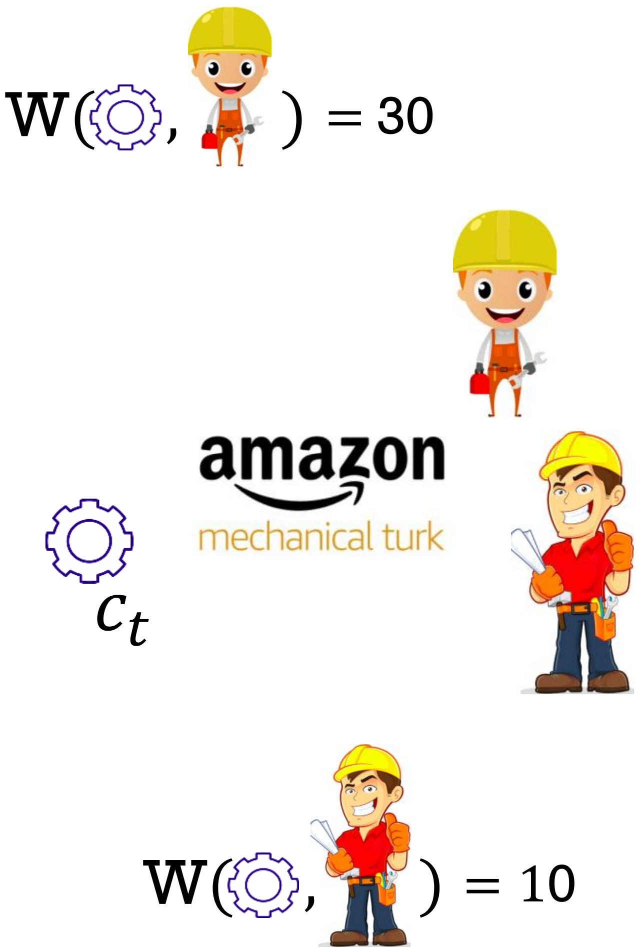
$$R(c_t, j) = < \theta_*, \phi(c_t, j) > + \eta_t$$

4. Costs are known:

$$W(c_t, j)$$

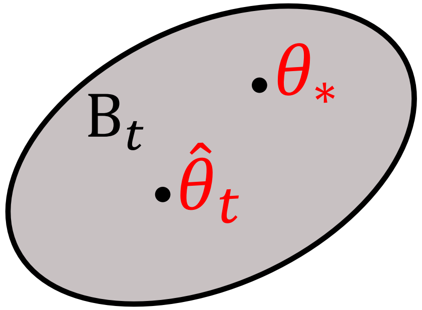
5. Stochastic (anytime) accumulative constraints:

$$E[\sum_{t=1}^{\tau} W(c_t, A(t))] \leq E[\sum_{t=1}^{\tau} U(t)]$$



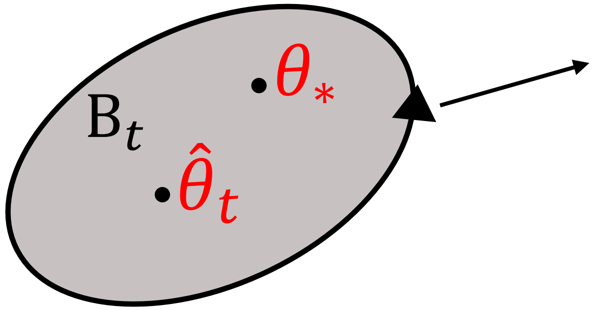
BANDIT LEARNING-BASED ONLINE ALGORITHM

1. Optimistic reward estimation

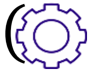


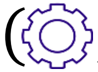
BANDIT LEARNING-BASED ONLINE ALGORITHM

1. Optimistic reward estimation



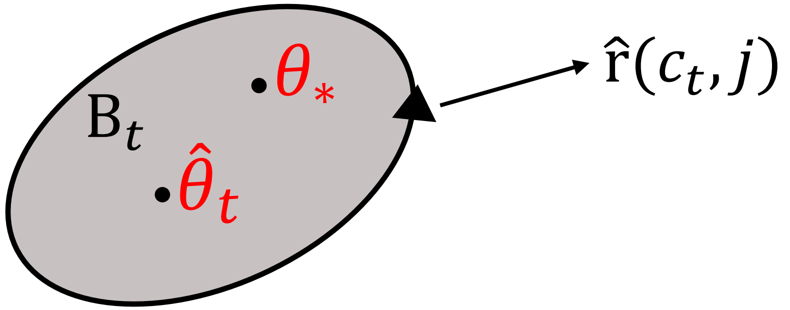
$\hat{\mathrm{r}}(c_t, j)$

$\widehat{\text{reward}}($  ,  $) = 0.9$

true reward( ,  $) = 0.8$

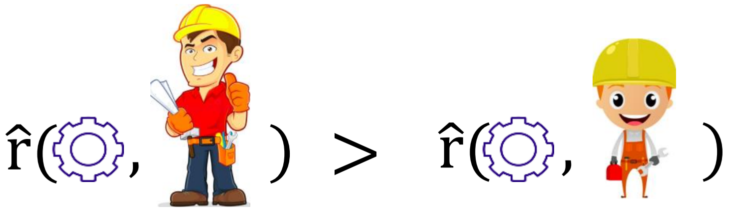
BANDIT LEARNING-BASED ONLINE ALGORITHM

1. Optimistic reward estimation



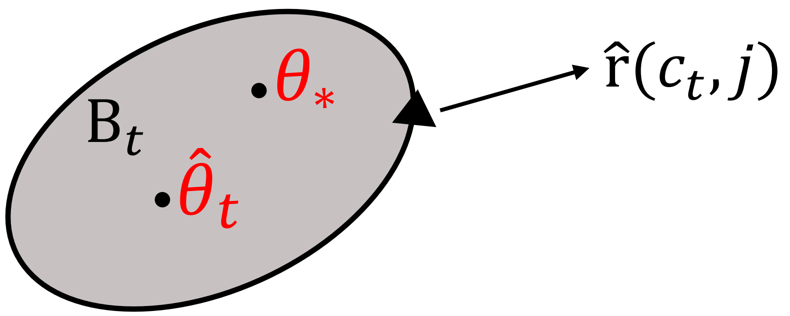
2. Pessimistic action

$$A(t) = \underset{j}{\operatorname{argmax}}\ \widehat{\text{reward}}(c_t, j) - \text{violation}(c_t, j)$$



BANDIT LEARNING-BASED ONLINE ALGORITHM

1. Optimistic reward estimation

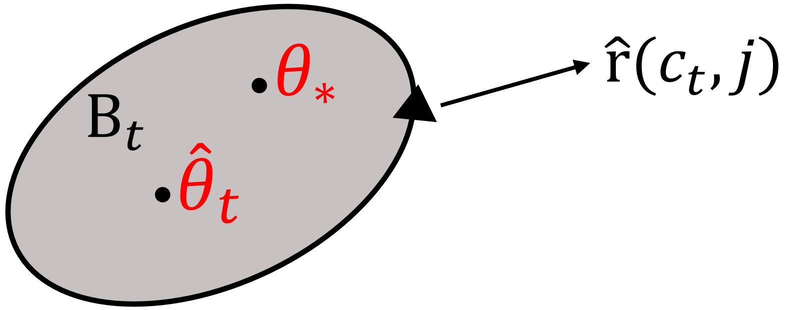


2. Pessimistic action

$$\mathrm{A(t)} = \underset{j}{\operatorname{argmax}}\ \widehat{\mathrm{reward}}(c_t, j) - \mathrm{violation}(c_t, j)$$

BANDIT LEARNING-BASED ONLINE ALGORITHM

1. Optimistic reward estimation



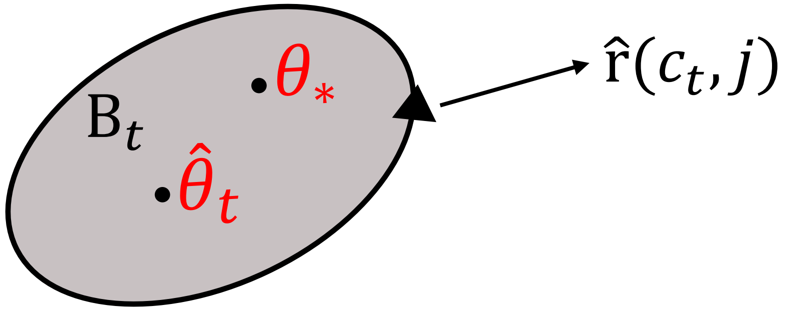
2. Pessimistic action

$$\mathrm{A(t)} = \underset{j}{\operatorname{argmax}}\ \widehat{\mathrm{reward}}(c_t, j) - \mathrm{violation}(c_t, j)$$

3. Calibration

BANDIT LEARNING-BASED ONLINE ALGORITHM

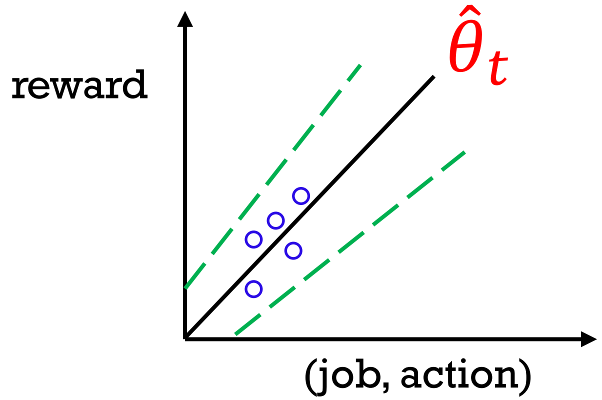
1. Optimistic reward estimation



2. Pessimistic action

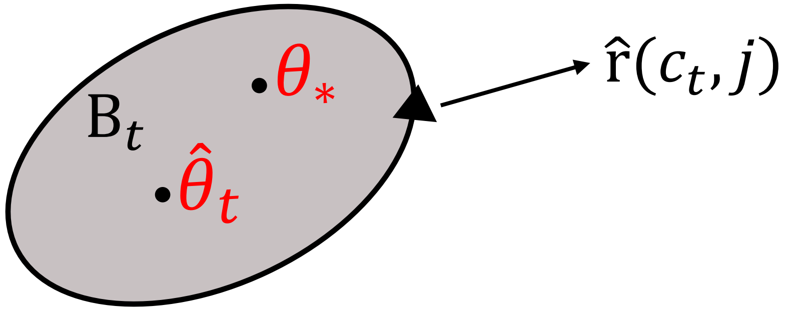
$$A(t) = \underset{j}{\operatorname{argmax}} \ \widehat{\text{reward}}(c_t, j) - \text{violation}(c_t, j)$$

3. Calibration on reward



BANDIT LEARNING-BASED ONLINE ALGORITHM

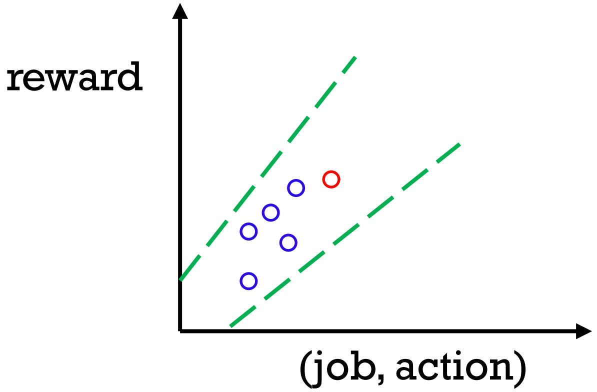
1. Optimistic reward estimation



2. Pessimistic action

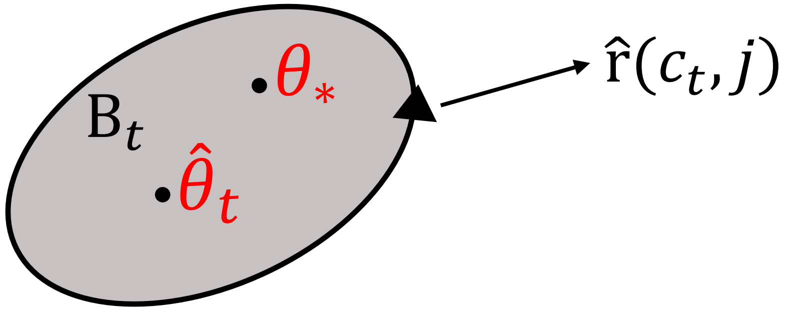
$$A(t) = \underset{j}{\operatorname{argmax}} \ \widehat{\text{reward}}(c_t, j) - \text{violation}(c_t, j)$$

3. Calibration on reward



BANDIT LEARNING-BASED ONLINE ALGORITHM

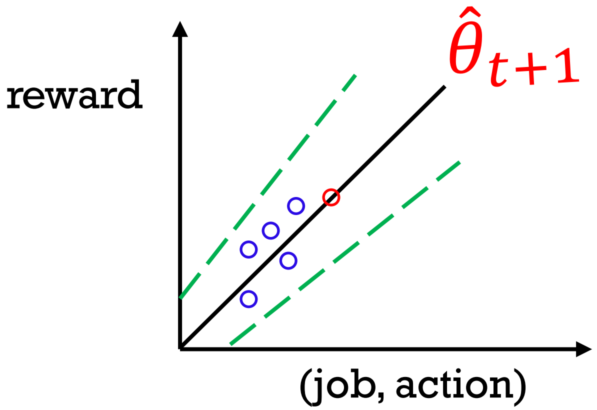
1. Optimistic reward estimation



2. Pessimistic action

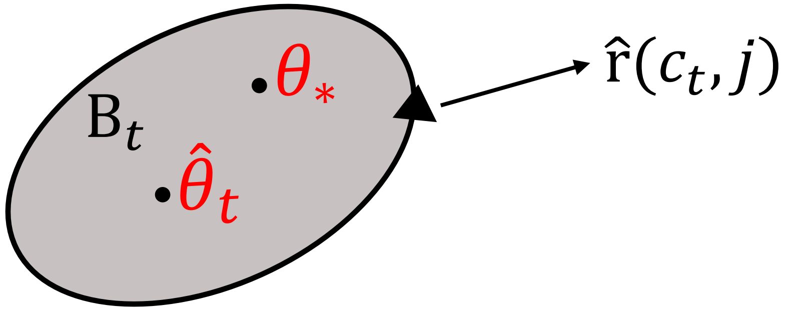
$$A(t) = \underset{j}{\operatorname{argmax}}\ \widehat{\text{reward}}(c_t, j) - \text{violation}(c_t, j)$$

3. Calibration on reward



BANDIT LEARNING-BASED ONLINE ALGORITHM

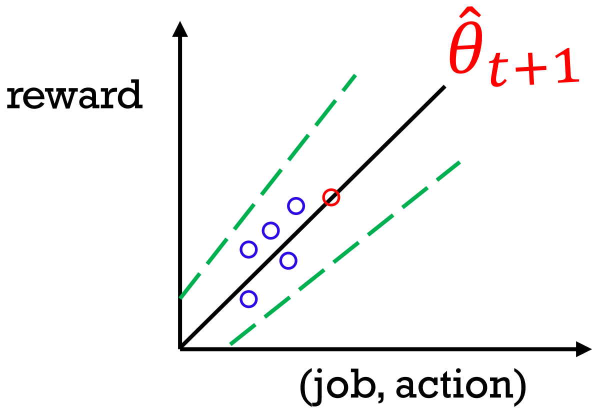
1. Optimistic reward estimation

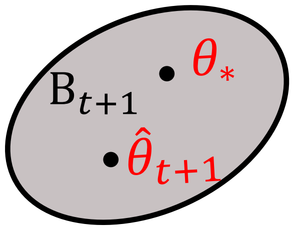


2. Pessimistic action

$$A(t) = \underset{j}{\text{argmax}}\ \widehat{\text{reward}}(c_t, j) - \text{violation}(c_t, j)$$

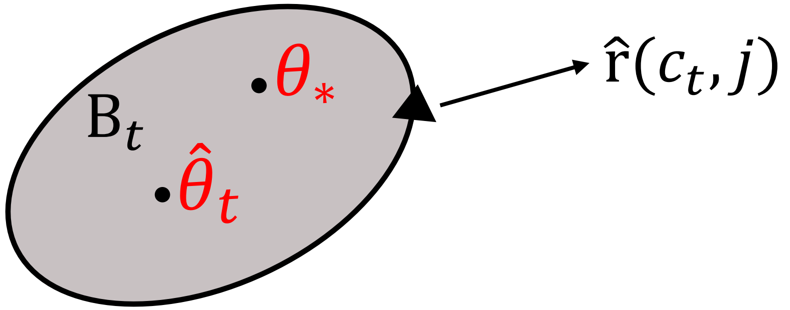
3. Calibration on reward





BANDIT LEARNING-BASED ONLINE ALGORITHM

1. Optimistic reward estimation

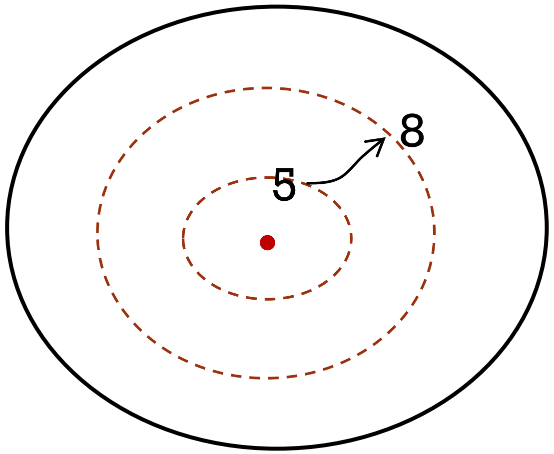


2. Pessimistic action

$$A(t) = \underset{j}{\operatorname{argmax}}\ \widehat{\text{reward}}(c_t, j) - \text{violation}(c_t, j)$$

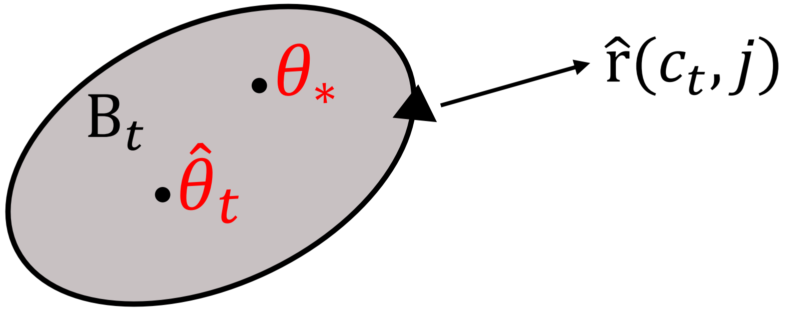
3. Calibration on violation

$$\text{violation}(t+1) = \underbrace{\text{violation}(t)}_{5} + \underbrace{\text{cost}(t) - \text{budget}(t)}_{3}$$



PESSIMISTIC-OPTIMISTIC ONLINE ALGORITHM

1. Optimistic reward estimation

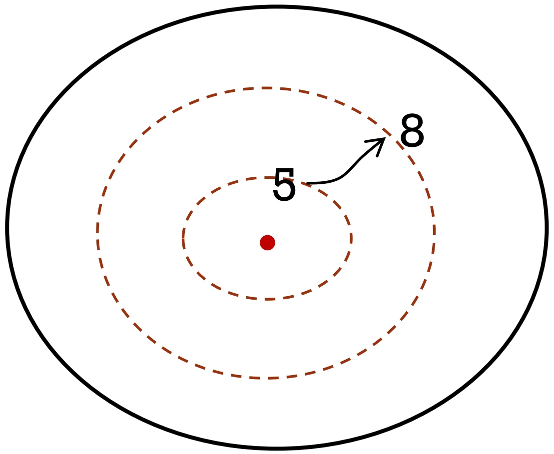


2. Pessimistic action

$$\mathrm{A(t)} = \underset{j}{\operatorname{argmax}}\ \widehat{\mathrm{reward}}(c_t, j) - \mathrm{violation}(c_t, j)$$

3. Calibration on violation

$$\mathrm{violation}(t+1) = \underbrace{\mathrm{violation}(t)}_{5} + \underbrace{\mathrm{cost}(t) - \mathrm{budget}(t)}_{3}$$



PESSIMISTIC-OPTIMISTIC ONLINE ALGORITHM

Theorem (Informal):

Pessimistic-optimistic algorithm achieves Regret(τ) = O($\sqrt{\tau}$) and Violation(τ) = 0 after some constant rounds.

[LiuLiShiYing21] First efficient online algorithm to achieve optimal regret & violation (anytime).

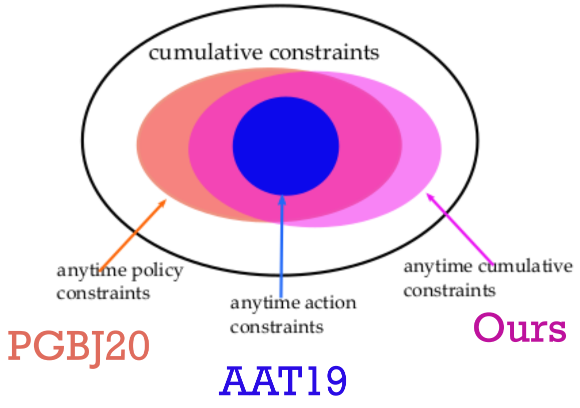
PESSIMISTIC-OPTIMISTIC ONLINE ALGORITHM

Theorem (Informal):

Pessimistic-optimistic algorithm achieves Regret(τ) = O($\sqrt{\tau}$) and Violation(τ) = 0 after some constant rounds.

[LiuLiShiYing21] First efficient online algorithm to achieve optimal regret & violation (anytime).

Related Work	**Constriant Type**
AD14, AD16, BKS18, CER20	Constraints imposed at the end of time horizon
AAT19	Anytime action constraints
PGBJ20	Anytime policy constraints



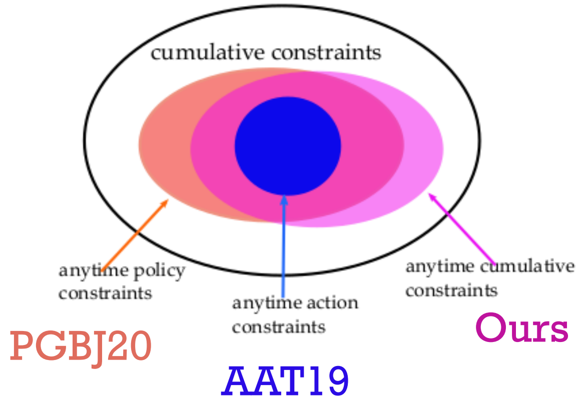
PESSIMISTIC-OPTIMISTIC ONLINE ALGORITHM

Theorem (Informal):

Pessimistic-optimistic algorithm achieves Regret(τ) = O($\sqrt{\tau}$) and Violation(τ) = 0 after some constant rounds.

[LiuLiShiYing21] First efficient online algorithm to achieve optimal regret & violation (anytime).

Related Work	Constraint Type
AD14, AD16, BKS18, CER20	Constraints imposed at the end of time horizon
AAT19	Anytime action constraints
PGBJ20	Anytime policy constraints



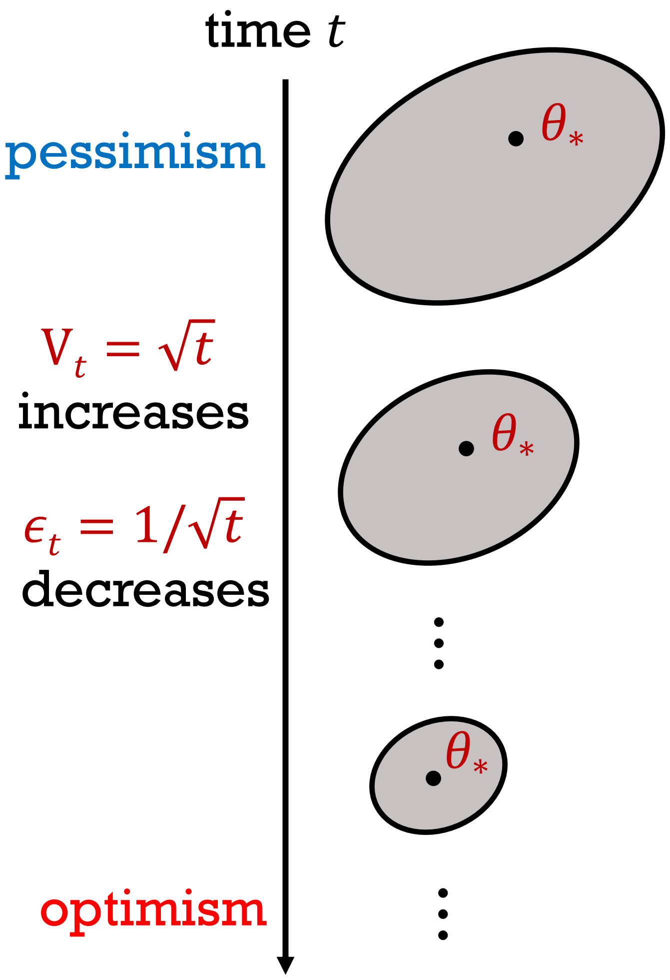
Primal-dual approach with adaptive optimism in primal and pessimism in dual.

ADAPTIVE OPTIMISM-PESSIMISM IN PRIMAL-DUAL

Primal (action):

$$A(t) = \underset{j}{\operatorname{argmax}} \ \ \widehat{\text{reward}}(c_t, j) - \text{violation}(c_t, j)$$

$$= \underset{j}{\operatorname{argmax}} \ \ \mathrm{V}_t \hat{\mathrm{r}}(c_t, j) - \text{violation}(c_t, j)$$



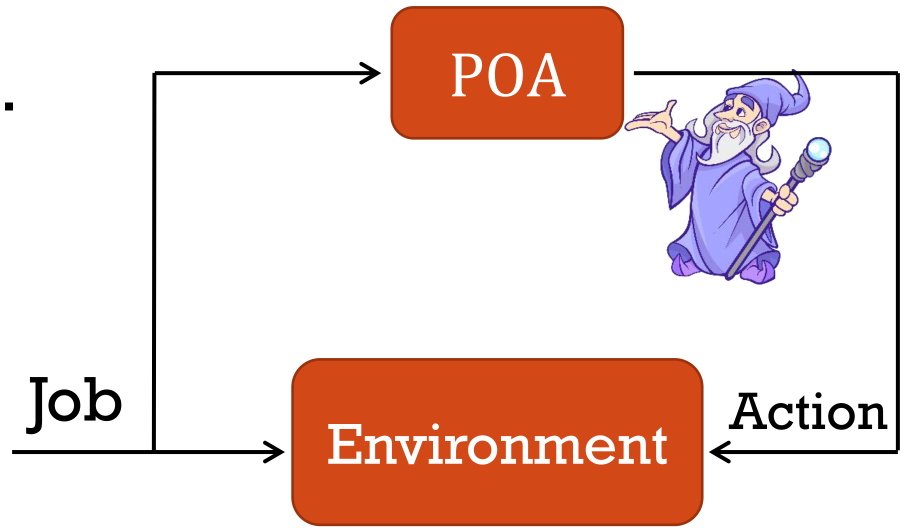
Dual (calibration):

$$\text{violation}(t+1) = \text{violation}(t) + \text{cost}(t) - \text{budget}(t) + \epsilon_t$$

CONCLUSION

Pessimistic-optimistic online algorithm:

- achieve optimal regret & violation (anytime).
- a novel drift analysis framework to bridge regret and violation.



THANK YOU!