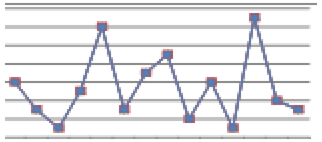
NREL, 7th Workshop on Autonomous Energy Systems

Online Learning for Residential Demand Response via Advanced Multi-Armed Bandits

Dr. Xin Chen

Assistant Professor
Electrical & Computer Engineering
Texas A&M University
Email: xin_chen@tamu.edu
09/03/2024

Residential Demand Response



Generation = **Demand** + Loss

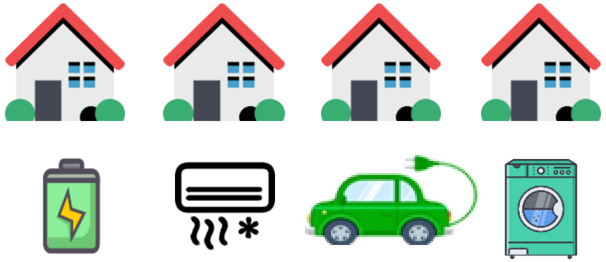
(adjustment)

Power System

Aggregator

incentive

load adjustment

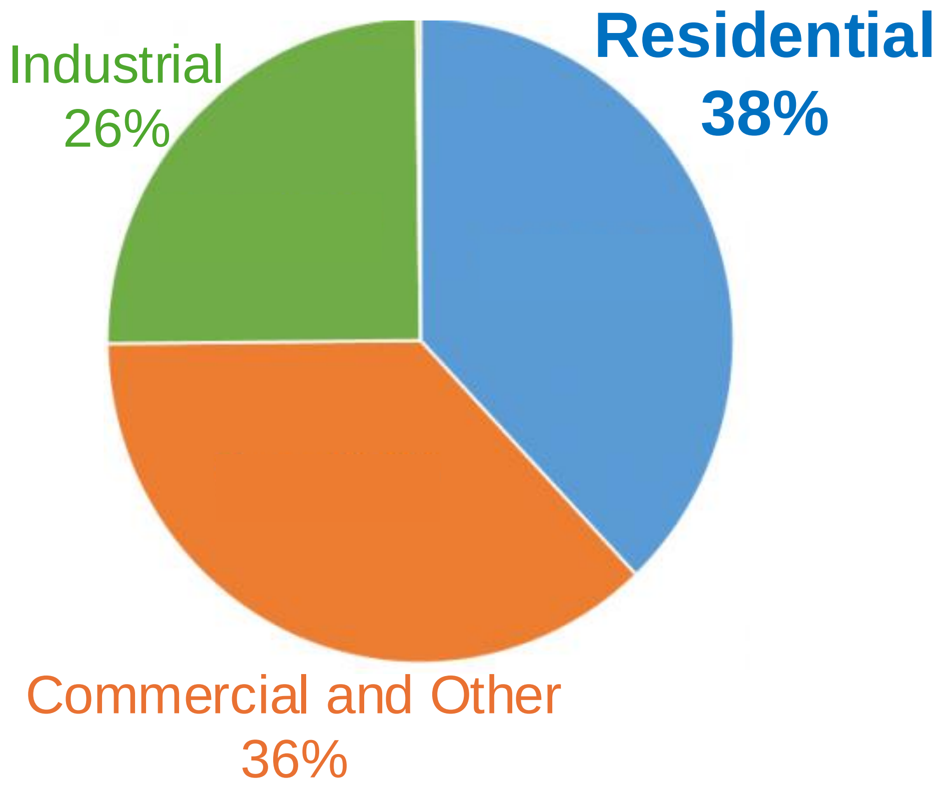


End-Users

Why Study Residential Demand Response?

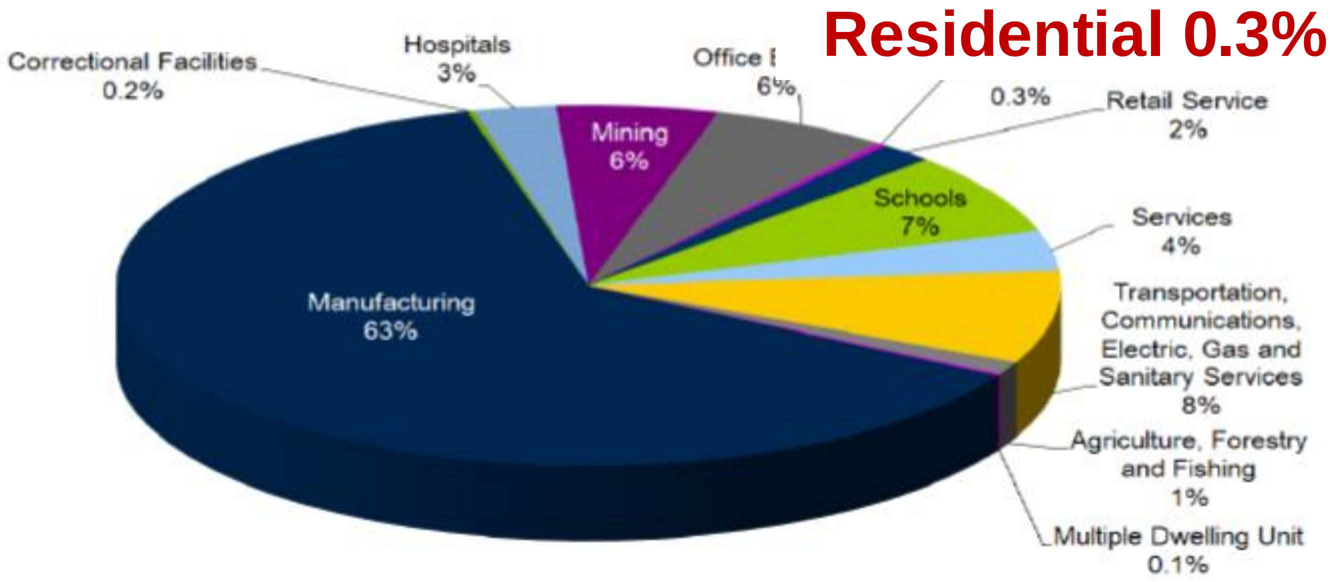
Because **residential** demand takes the **largest share**;

Huge potential but **underutilized**.



Source: 2022 U.S. Energy Information Administration (EIA), “Annual Electric Power Industry Report”.

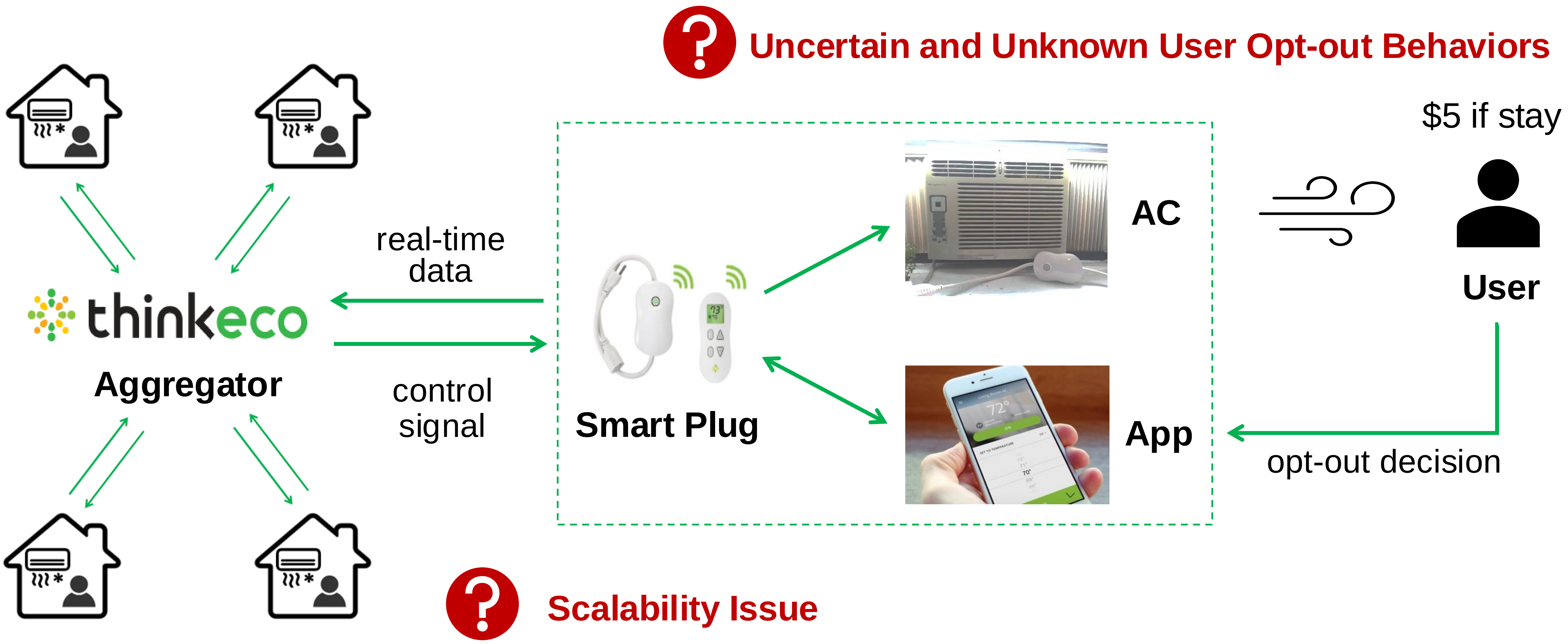
20/21 Confirmed DR registrations business segments



Source: PJM, “Demand response operations market activity report,” 2022.
Percent of load capacity (MWs)

Goal: Select Right Users to Signal

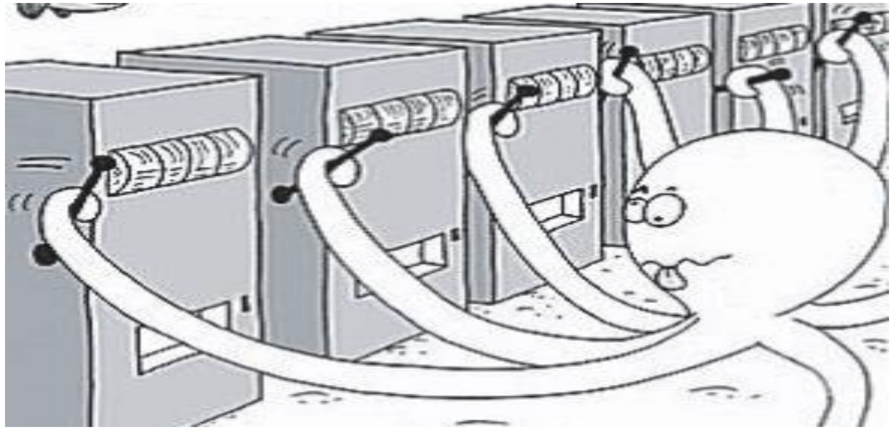
Due to budget constraints, need to select a subset of users (e.g., 1k) from the user pool (e.g., 10k)



Multi-Armed Bandits (MAB) Framework

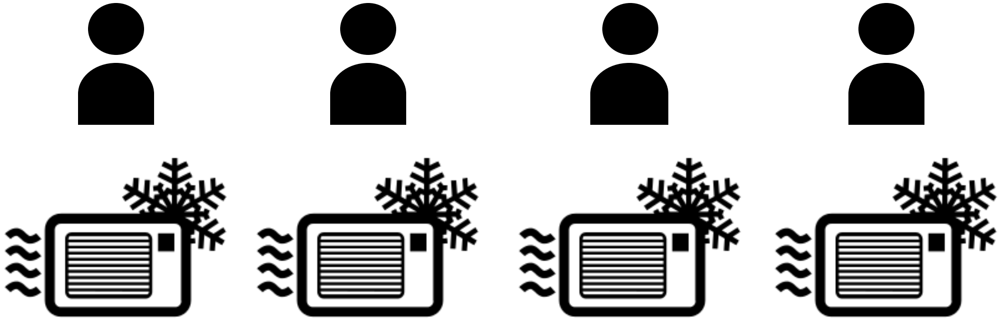
❖ Key Features: 1) Arms are different but independent; 2) Uncertain and unknown behaviors

Bandit Slot Machine



- Select one arm to maximize the profits;
- Observe the reward of the selected arm;
- Improve play strategies from feedback.

Demand Response



- Select a subset of users for DR;
- Observe responses from selected users;
- Learn users’ behaviors from responses.

Application Examples

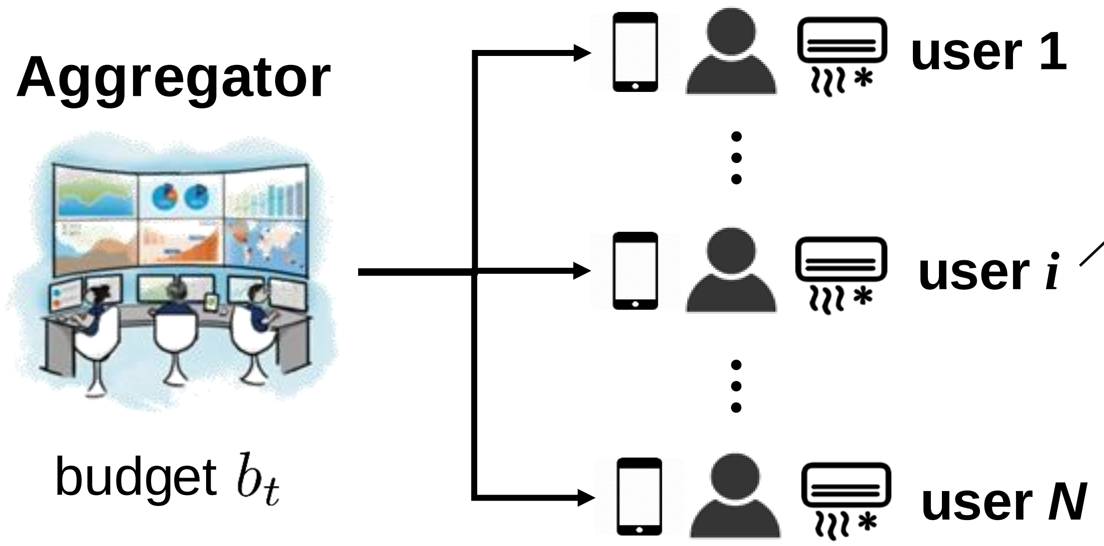
EV charging management

Residential load control

Problem Formulation

Consider a time horizon $[T] = \{1, 2, \cdots, T\}$

Each time $t \in [T]$ denotes a DR event.



- Binary random variable $z_{i,t} \in \mathrm{Bern}(\boxed{p_{i,t}})$.

$$z_{i,t} = \begin{cases} 1 & \text{: Stay-in} \\ 0 & \text{: Opt-out} \end{cases}$$

Unknown

- Adjustable load capacity $c_{i,t}$.
- Incentive credit $r_{i,t}$.

Optimal User Selection Model

$$\text{Obj.} \quad \max_{\mathcal{S}_t \subseteq [N]} \mathbb{E}\Big(\sum_{i \in \mathcal{S}_t} c_{i,t} z_{i,t}\Big) = \sum_{i \in \mathcal{S}_t} c_{i,t} \boxed{p_{i,t}}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{S}_t} r_{i,t} \leq b_t$$

Unknown

Select a subset $\mathcal{S}_t$ of users to maximize expected total load reduction.

budget constraint

Solution: Contextual Multi-Armed Bandits

❖ **Logistic regression** to predict $p_{i,t}$ for each user *i* at time *t under contextual influence*:

$$p_{i,t} = g(\boldsymbol{\theta}_i^\top \boldsymbol{x}_{i,t}) = \frac{1}{1 + \exp\left(-\boldsymbol{\theta}_i^\top \boldsymbol{x}_{i,t}\right)}$$

(Unknown) Individual Preference

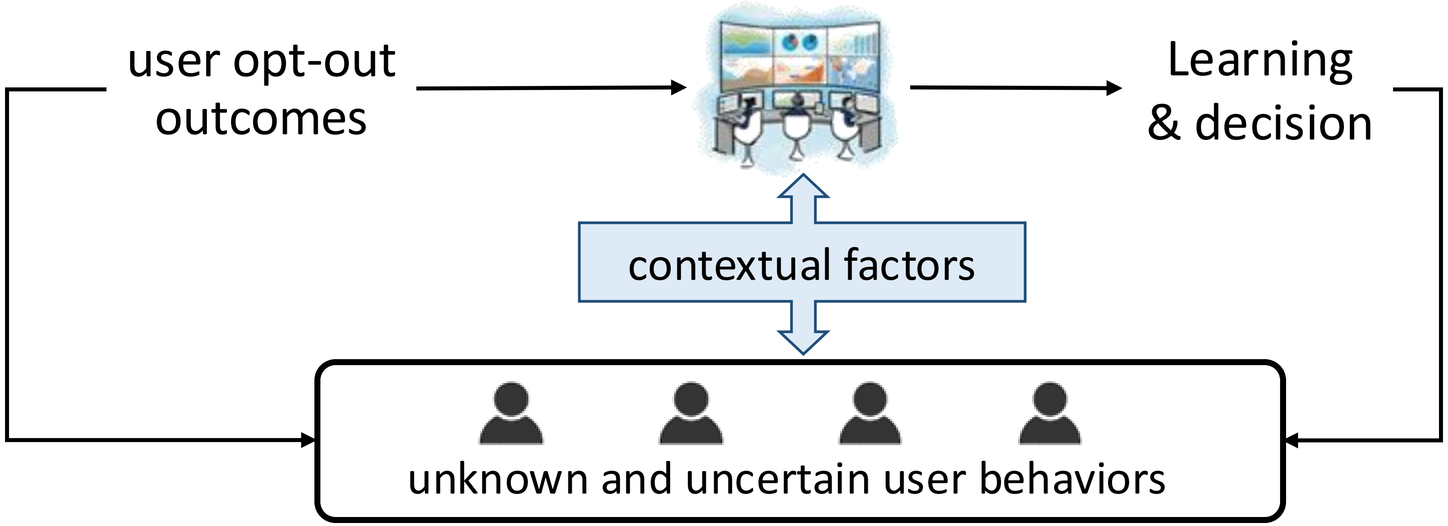
$$\boldsymbol{\theta}_i = \left(\theta_i^{(1)}, \theta_i^{(2)}, \cdots, \theta_i^{(m)}\right)$$

Contextual Factors

$$\boldsymbol{x}_{i,t} = \left(1, x_{i,t}^{(1)}, \cdots, x_{i,t}^{(m)}\right)$$

(electricity price, credit, weather, temperature, …)

❖ **Online Learning and Human-In-the-Loop Decision:**



Thompson Sampling to learn unknown $\boldsymbol{\theta}_i$ with balance of **exploitation** and **exploration**

Online Algorithm Based on Thompson Sampling.

- Assume unknown $\boldsymbol{\theta}_i$ be a random variable with Gaussian prior $\mathbb{P}_{\boldsymbol{\theta}_i} = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
- In each demand response event t,

$$p_{i,t} = \frac{1}{1 + \exp\left(-\hat{\boldsymbol{\theta}}_i^\top \boldsymbol{x}_{i,t}\right)}$$

Step 1: Sample $\hat{\boldsymbol{\theta}}_i$ from its distribution $\mathbb{P}_{\boldsymbol{\theta}_i}$.

Step 2: Select users by solving

$$\begin{aligned} \text{Obj.} \quad & \max_{\mathcal{S}_t \subseteq [N]} \mathbb{E}\Big(\sum_{i \in \mathcal{S}_t} c_{i,t} z_{i,t}\Big) = \sum_{i \in \mathcal{S}_t} c_{i,t} p_{i,t} \\ \text{s.t.} \quad & \sum_{i \in \mathcal{S}_t} r_{i,t} \le b_t \end{aligned}$$

$$\begin{aligned} \text{Obj.} \quad & \max_{\alpha_{i,t} \in \{0,1\}} \sum_{i=1}^{N} c_{i,t} p_{i,t} \alpha_{i,t} \\ \text{s.t.} \quad & \sum^{N} r_{i,t} \alpha_{i,t} \le b_t \end{aligned}$$

Binary Optimization

Step 3: Update posterior $\mathbb{P}_{\boldsymbol{\theta}_i} \leftarrow \mathbb{P}_{\boldsymbol{\theta}_i}(\cdot | \boldsymbol{x}_{i,t}, z_{i,t})$ with the observation $\boldsymbol{x}_{i,t}, z_{i,t}$.

variational Bayesian inference approach [3]

[3] T. S. Jaakkola and M. I. Jordan, “A variational approach to Bayesian logistic regression models and their extensions”. in Sixth International Workshop on Artificial Intelligence and Statistics, vol. 82, pp. 4, 1997.

Regret Analysis

- T-time Regret: $\mathrm{Regret}(T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbb{E}\left[f_{\boldsymbol{\theta}}(\mathcal{S}_t^*, t) - f_{\boldsymbol{\theta}}(\mathcal{S}_t, t) \,|\, \boldsymbol{\theta}\right]$

  Optimal objective with true $\boldsymbol{\theta}$.

  Objective using the proposed algorithm.

- T-time Bayesian Regret: $\mathrm{BayesRegret}(T) = \mathbb{E}_{\boldsymbol{\theta} \sim P_0}\left[\mathrm{Regret}(T, \boldsymbol{\theta})\right]$

Theorem (informal): When T is sufficiently large, the Bayesian regret is

$$\mathrm{BayesRegret}(T) \leq O\left(N^2 \gamma^d \sqrt{T \log T (d + \log T)}\right) \sim \underline{O(\log(T)\sqrt{T})}$$

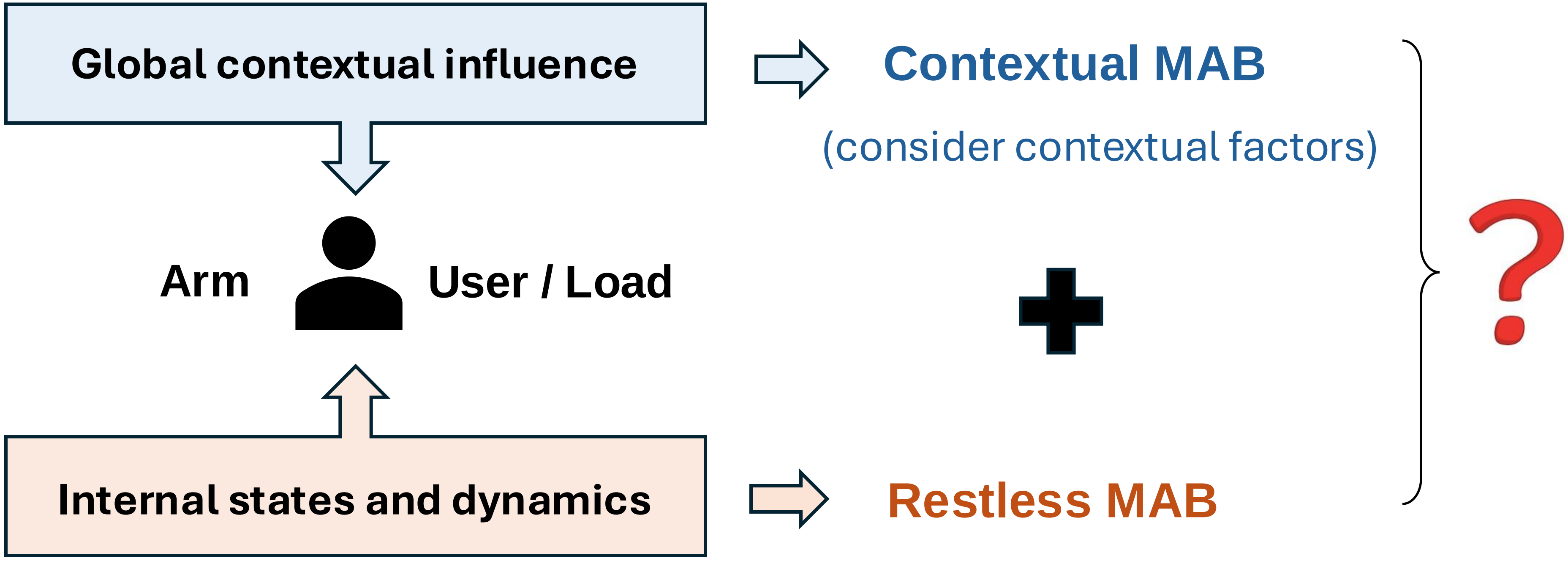
sublinear

where $\gamma = \exp(2 \sup_{i \in [N]} ||\boldsymbol{\theta}_i||_\infty)$ and d is the dimension of $\boldsymbol{\theta}_i$.

X. Chen, Y. Nie, and N. Li, "Online residential demand response via contextual multi-armed bandits," IEEE Control Systems Letters, 2020.

Recent Extension: Contextual Restless Bandits

- temperature, weather, price, time . . .



- Physical dynamics: e.g., AC thermal dynamics, EV charging SOC dynamics
- User “fatigue effect” . . .

Recent Extension: Contextual Restless Bandits

- temperature, weather, price, time . . .

Global contextual influence

Arm **User / Load**

Internal states and dynamics

- Physical dynamics: e.g., AC thermal dynamics, EV charging SOC dynamics
- User "fatigue effect" . . .

Contextual MAB

(consider contextual factors)

+

Restless MAB

(model each arm dynamics via MDP)

Our Novel Framework

Contextual Restless MAB (CRB)

- **X. Chen**, I. Hou, "Contextual Restless Multi-Armed Bandits with Application to Demand Response Decision-Making", IEEE CDC, 2024.

CRB Problem Formulation

Consider N arms and an infinite time horizon $t = 0, 1, 2, \cdots$

At each time t, observe the states $s_{i,t}$ of all arms and global context g_t, selects at most C_{g_t} arms

Decision-maker

find optimal selection policy $\pi(\boldsymbol{a}|\boldsymbol{s}, g)$

$$(\textbf{Primal}): \quad \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\sum_{i=1}^{N} \beta^t r_{i,t}\right]$$

$$\text{s.t.} \quad \sum_{i=1}^{N} a_{i,t} \le C_{g_t}, \quad \forall t \in \mathcal{T}$$

arm 1

⋮

arm *i*

⋮

arm N

Each arm i : **Context-augmented MDP**

- **Action**: $a_{i,t} \in \mathcal{A} := \{0, 1\}$
- **State**: $s_{i,t} \in \mathcal{S}$
- **Transition** (context-dependent):

  $P_i(s'|g, s, a) :=$

  $\mathbb{P}(s_{i,t+1} = s'|g_t = g, s_{i,t} = s, a_{i,t} = a)$
- **Reward** (context-dependent):

  $r_{i,t} = R_i(g_t, s_{i,t}, a_{i,t})$

- **Global context**: $g_t \in \mathcal{G}$

  follows positive-recurrent Markov Process

  $G(g'|g) := \mathbb{P}(g_{t+1} = g'|g_t = g)$

Solution Algorithm: Dual Decomposition

Very high-dimension MDP

$$(\textbf{Primal}): \quad \max_{\pi} \mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty}\sum_{i=1}^{N}\beta^t r_{i,t}\Big]$$

$$\text{s.t.} \quad \underline{\sum_{i=1}^{N} a_{i,t} \le C_{g_t},} \quad \forall t \in \mathcal{T}$$

Challenge ?

Relax budget constraint

expected per-global-context budget constraint

$$(\textbf{Relaxed}): \quad \max_{\pi} \mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty}\sum_{i=1}^{N}\beta^t r_{i,t}\Big]$$

$$\text{s.t.} \quad \mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty}\beta^t \mathbb{I}(g_t = g)\big(\sum_{i=1}^{N} a_{i,t}\big)\Big] \le \mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty}\beta^t \mathbb{I}(g_t = g)\Big] C_g, \quad \forall g \in \mathcal{G}$$

Lagrange multiplier $\boldsymbol{\lambda} := (\lambda_g)_{g\in\mathcal{G}}$

$$(\textbf{Dual}): \quad \min_{\boldsymbol{\lambda}\ge \mathbf{0}} \Big[\max_{\pi} L(\pi, \boldsymbol{\lambda})\Big].$$

$$L(\pi, \boldsymbol{\lambda}) = \sum_{i=1}^{N}\mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty}\beta^t (r_{i,t} - \lambda_{g_t} a_{i,t})\Big] + \mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty}\beta^t \lambda_{g_t} C_{g_t}\Big].$$

Dual Decomposition Solution

- Gradient descent update of $\boldsymbol{\lambda}$
- Each arm i solves a **local** problem

$$(\textbf{Arm}_i(\boldsymbol{\lambda})): \quad \max_{\rho_i} \mathbb{E}_{\rho_i}\Big[\sum_{t=0}^{\infty}\beta^t (r_{i,t} - \lambda_{g_t} a_{i,t})\Big]$$

Solution Algorithm: Dual Decomposition

- **Gradient descent update of $\boldsymbol{\lambda}$**

$$\lambda_g^{(k+1)} = \Big[\lambda_g^{(k)} + \delta_k \mathbb{E}_{\pi^*(\boldsymbol{\lambda}^{(k)})}\big[\sum_{t=0}^{\infty} \beta^t \mathbb{I}(g_t = g)\big(\sum_{i=1}^{N} a_{i,t} - C_g\big)\big]\Big]^+$$

$\pi^*(\boldsymbol{\lambda})$ $\qquad$ $\boldsymbol{\lambda}^{(k)} := (\lambda_g^{(k)})_{g\in\mathcal{G}}$

- Each arm *i* solves a **local unconstrained** problem

$$(\mathbf{Arm}_i(\boldsymbol{\lambda})): \ \max_{\rho_i} \mathbb{E}_{\rho_i}\Big[\sum_{t=0}^{\infty} \beta^t (r_{i,t} - \lambda_{g_t} a_{i,t})\Big]$$

the policy for arm *i*: $\rho_i(a|g, s, \boldsymbol{\lambda})$

Interpretation of λ: global “price” paid to select an arm

Solve Bellman Optimality Equation

✓ Optimal Q-function: $Q^*_{i,\boldsymbol{\lambda}}(g, s, a) = R_i(g, s, a) - \lambda_g a + \beta \sum_{g'\in\mathcal{G}, s'\in\mathcal{S}} G(g'|g) P_i(s'|g, s, a) V^*_{i,\boldsymbol{\lambda}}(g', s').$

Optimal local policy:

$$\rho_i^*(\boldsymbol{\lambda}): a_{i,t} = \begin{cases} 1, & \text{if } Q^*_{i,\boldsymbol{\lambda}}(g_t, s_{i,t}, 1) > Q^*_{i,\boldsymbol{\lambda}}(g_t, s_{i,t}, 0) \\ 0, & \text{otherwise.} \end{cases}$$

Index Policy with Known Models

Real-time application:

1. Compute index of each arm
2. Rank arms by index values
3. Select top C_{g_t} arms

Algorithm 1 Index Policy Algorithm for Solving the CRB Problem with Known Arm Models.

Initialization: $\boldsymbol{\lambda}^{(0)} := (\lambda_g^{(0)})_{g \in \mathcal{G}} \leftarrow \mathbf{0}$; $k \leftarrow 0$; $\epsilon > 0$.
repeat
  Perform the following three steps for each arm $i \in [N]$ in parallel:
    1) Solve $\mathbf{LP}_i(\boldsymbol{\lambda}^{(k)})$ (9) to obtain $V^*_{i,\boldsymbol{\lambda}^{(k)}}(g,s)$;
    2) Compute the Q-function $Q^*_{i,\boldsymbol{\lambda}^{(k)}}(g,s,a)$ by (10);
    3) Construct the optimal policy $\rho_i^*(\boldsymbol{\lambda}^{(k)})$ by (11).
  Let $\pi^*(\boldsymbol{\lambda}^{(k)}) := (\rho_i^*(\boldsymbol{\lambda}^{(k)}))_{i \in [N]}$, and perform the update (7) to obtain $\boldsymbol{\lambda}^{(k+1)}$, where the expectation is computed using (12)-(14). Let $k \leftarrow k+1$.
until the convergence $||\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k-1)}|| \leq \epsilon$ is met.
Let $\boldsymbol{\lambda}^* \leftarrow \boldsymbol{\lambda}^{(k)}$.
for time $t = 0, 1, 2, \ldots$ **do**
  Compute the index of each arm $i \in [N]$:

$$I_{i,t} \leftarrow Q^*_{i,\boldsymbol{\lambda}^*}(g_t, s_{i,t}, 1) - Q^*_{i,\boldsymbol{\lambda}^*}(g_t, s_{i,t}, 0). \quad (15)$$

  Sort all arms such that $I_{(1),t} \geq I_{(2),t} \geq \cdots \geq I_{(N),t}$.
  Activate the top C_{g_t} arms $(1), (2), \cdots, (C_{g_t})$.
end for

Index of each arm

Solve Relaxed Problem to obtain $\boldsymbol{\lambda}^*$ via dual decomposition

Implement Index Policy (**ensure budget constraint**)

Online Learning with Unknown Models

Algorithm 2 Online Learning Algorithm for Solving the CRB Problem with Unknown Arm Models.

1: **Initialization**: Time window T; $\epsilon_n > 0$; initial transition kernel $P_i^0(\cdot)$ of each arm $i \in [N]$.
2: **for** epoch $n = 0, 1, 2, \ldots$ **do**
3: Based on the up-to-date transition kernel model $P_i^n(\cdot)$ of each arm $i \in [N]$, follow Steps 2-6 in Algorithm 1 to compute the optimal $\boldsymbol{\lambda}_n^*$.
4: **for** time $t = nT, nT+1, \cdots, nT+T-1$ **do**
5: With probability of $1 - \epsilon_n$,
- Compute the index $I_{i,t}$ (16) of each arm $i \in [N]$;
- Sort all arms such that $I_{(1),t} \geq I_{(2),t} \geq \cdots \geq I_{(N),t}$ and activate the top C_{g_t} arms.

With probability of ϵ_n, randomly activate C_{g_t} arms.
6: **end for**
7: For each arm $i \in [N]$, based on the observed state transitions, update its transition kernel model by:

$$P_i^{n+1}(s_{i,t+1} = s' | g_t = g, s_{i,t} = s, a_{i,t} = a) = \frac{M^i_{s',g,s,a}}{M^i_{g,s,a}},$$

where $M^i_{g,s,a}$ and $M^i_{s',g,s,a}$ are arm i's cumulative historical counts of the context-state-action tuple (g, s, a) and the state transition $(g, s, a) \to s'$.
8: **end for**

Solve Relaxed Problem

Implement Index Policy
with ϵ-exploration

Better online learning algorithms:
Thompson Sampling, UCB, Q-learning

Update estimation of transition kernel models

Function approximation for Scalability

Theoretical Analysis: Asymptotical Optimality

Optimal Objective Values: $V_{\text{Rel}}^N \geq V_{\text{Pri}}^N \geq V_{\text{Ind}}^N$

Relaxed problem

Primary problem

Index policy
(feasible but maybe sub-optimal)

Theorem 1. *Suppose that the initial global context g_0 is chosen uniformly at random from $\mathcal{G}$ and the initial state $s_{i,0}$ of each arm $i \in [N]$ is chosen independently with the distribution $\mathbb{P}(s_{i,0} = s | g_0 = g) = m_g^*(s)$, then, under Assumption 1, we have*

$$V_{\text{Rel}}^N \geq V_{\text{Pri}}^N \geq V_{\text{Rel}}^N - \mathcal{O}(\sqrt{N}). \tag{20}$$

as $N \to \infty$

$$V_{\text{Rel}}^N / N \longrightarrow V_{\text{Pri}}^N / N$$

X. Chen, I. Hou, “Contextual Restless Multi-Armed Bandits with Application to Demand Response Decision-Making“, IEEE CDC, 2024.

Numerical Simulations

Residential Demand Response:

- Number of users (arms) $N = 500$
- Discrete global context

$$g \in \mathcal{G} = \{1, 2, \cdots, 6\}$$

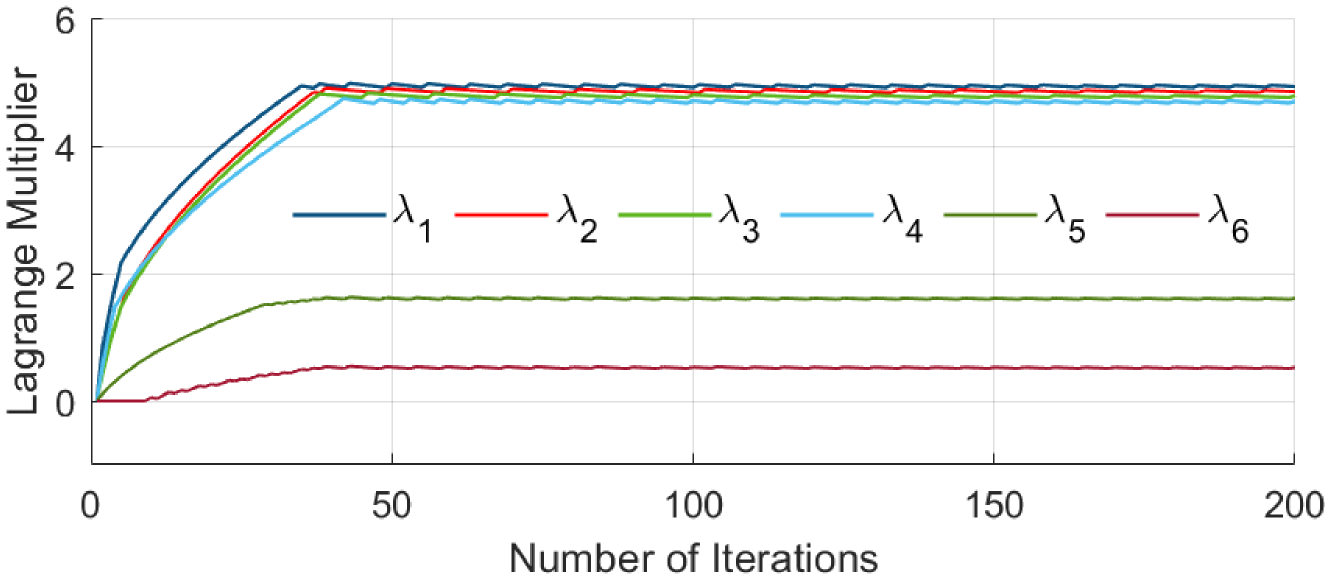
- Selection budget C = 100
- State: $s_{i,t} := (z_{i,t}, x_{i,t})$

$$x_{i,t} \in \{1, 2, \cdots, 4\}$$

- Reward: $R_i(\cdot) = \frac{a_{i,t} z_{i,t} l_i}{(g_t - x_{i,t})^2 + 1}$
- Discount factor $\beta = 0.97$

finite time horizon $T = 300$

Fig 1. Convergence of Dual Decomposition Alg.



Numerical Simulations

$$V_{\mathrm{Rel}}^N \geq V_{\mathrm{Pri}}^N \geq V_{\mathrm{Ind}}^N$$

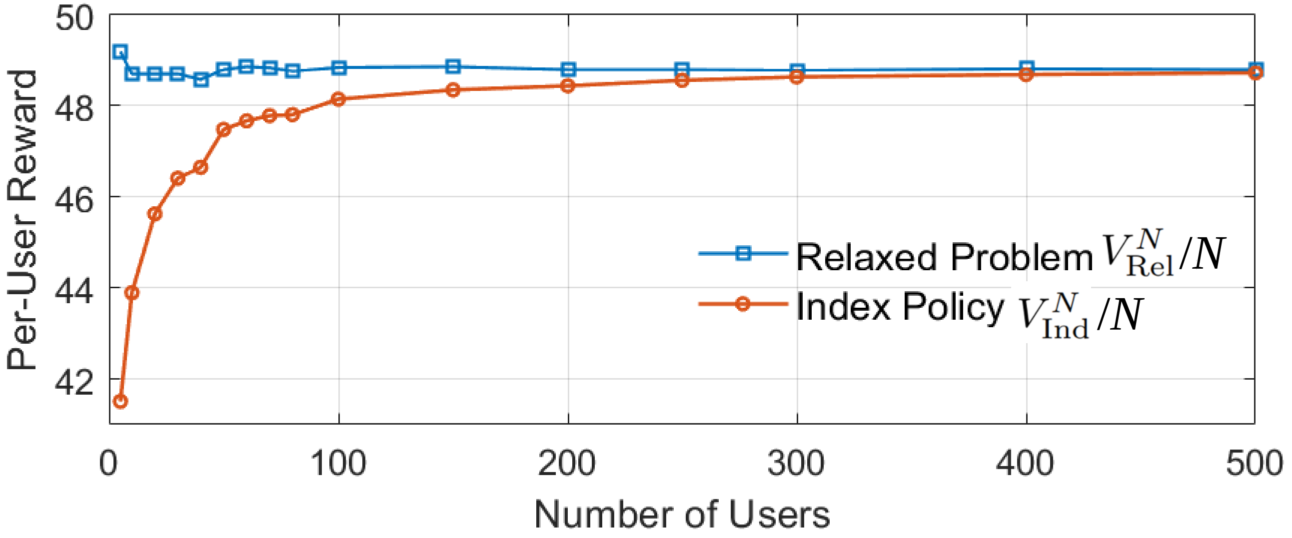


Fig 2. Asymptotical optimality of index policy.

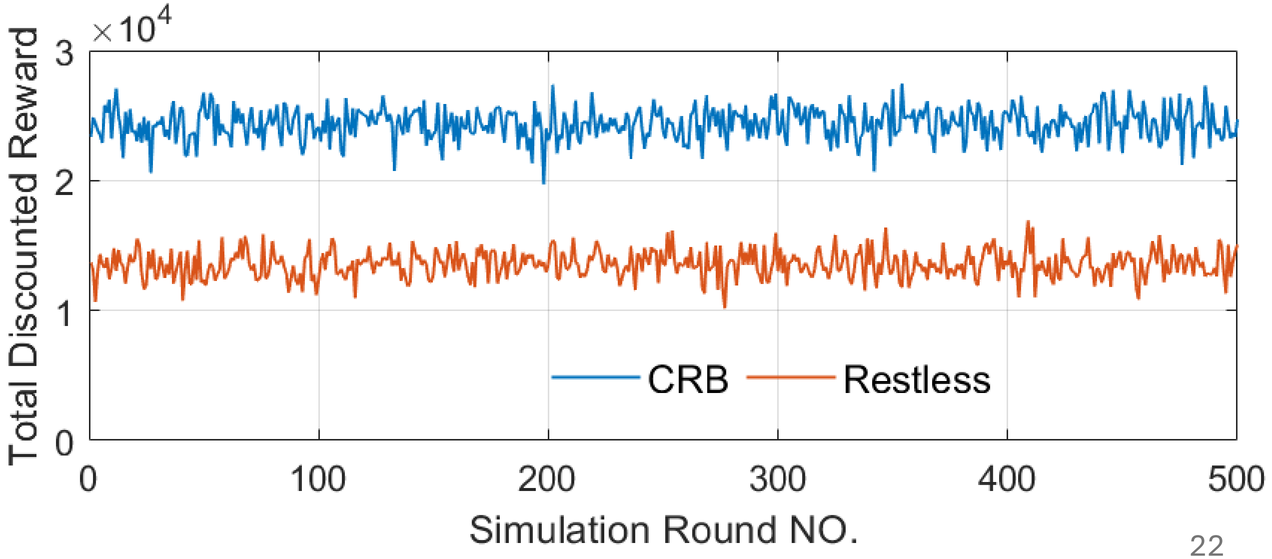
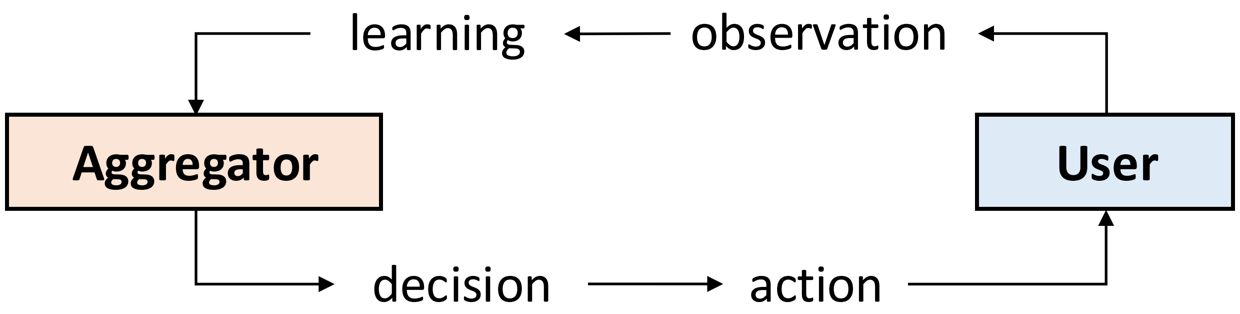


Fig 3. Comparison with Restless bandits.

Key Takeaway

Online learning for **human-in-the-loop DR decision**



- The **multi-armed bandits (MAB)** method is useful for large-scale DR online decision-making.
- A novel MAB framework, **Contextual Restless Bandits (CRB),** models both the **dynamic state transitions of each arm** and the influence of **external global environmental context.**
- A scalable **index policy** algorithm based on dual decomposition is proposed to solve CRB.
- Simulation results demonstrate the **asymptotic optimality** and enhanced **modeling capability**.

References

[1] A. Slivkins et al., "Introduction to multi-armed bandits," Foundations and Trends in Machine Learning, vol. 12, no. 1-2, pp. 1–286, 2019

[2] L. Zhou, "A survey on contextual multi-armed bandits," arXiv preprint, arXiv:1508.03326, 2015.

[3] P. Whittle, "Restless bandits: Activity allocation in a changing world," Journal of Applied Probability, vol. 25, no. A, pp. 287–298, 1988.

[4] R. R. Weber and G. Weiss, "On an index policy for restless bandits," Journal of Applied Probability, vol. 27, no. 3, pp. 637–648, 1990.

[5] X. Chen, I. Hou, "Contextual Restless Multi-Armed Bandits with Application to Demand Response Decision-Making", arXiv:2403.15640, 2024.

[6] X. Chen, Y. Li, J. Shimada, and N. Li, "Online learning and distributed control for residential demand response," IEEE Transactions on Smart Grid, vol. 12, no. 6, pp. 4843–4853, 2021.

[7] X. Chen, Y. Nie, and N. Li, "Online residential demand response via contextual multi-armed bandits," IEEE Control Systems Letters, vol. 5, no. 2, pp. 433–438, 2020.

[8] B. Liang, L. Xu, A. Taneja, M. Tambe, and L. Janson, "A Bayesian approach to online learning for contextual restless bandits with applications to public health," arXiv preprint arXiv:2402.04933, 2024

[9] J. A. Taylor and J. L. Mathieu, "Index policies for demand response," IEEE Transactions on Power Systems, vol. 29, no. 3, pp. 1287–1295, 2013.

[10] N. Gast, B. Gaujal, and C. Yan, "Exponential convergence rate for the asymptotic optimality of whittle index policy," arXiv preprint arXiv:2012.09064, 2020.