THE WEB CONFERENCE ACM

The Grainger College of Engineering
IBM-Illinois Discovery Accelerator Institute

AIFARMS
Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

NSF

Neural Contextual Bandits for Personalized Recommendation

Yikun Ban **Yunzhe Qi** **Jingrui He**

University of Illinois Urbana-Champaign

{Yikunb2, Yunzheq2, Jingrui}@illinois.edu

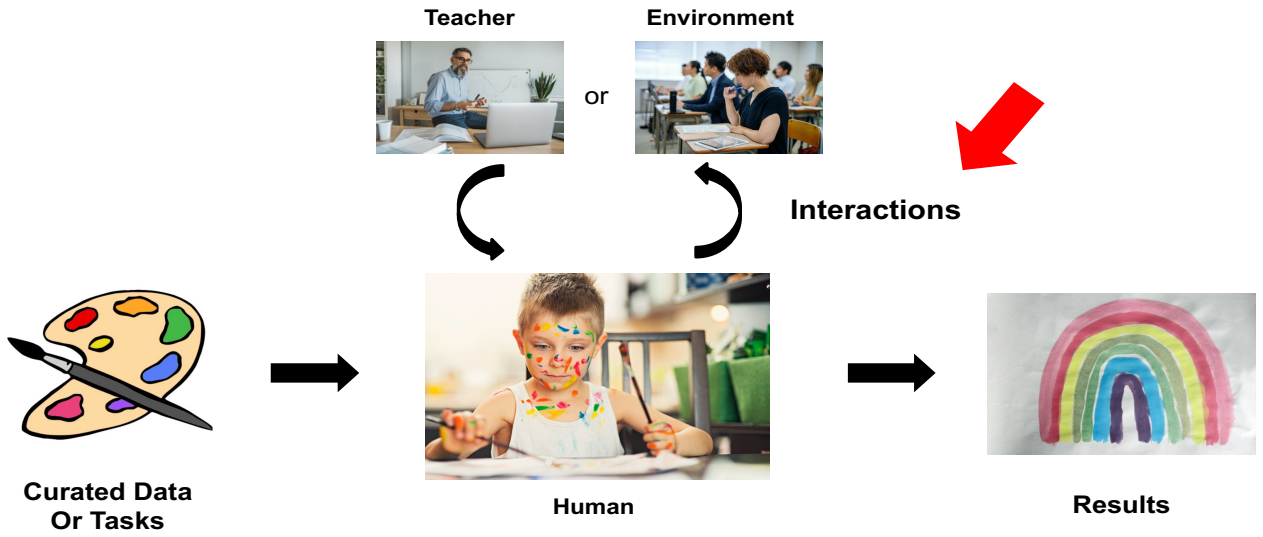
Time: 9:00 AM – 12:30 PM, 13 May 2024

Location: Virgo 1, Resorts World Sentosa Convention Centre, Singapore

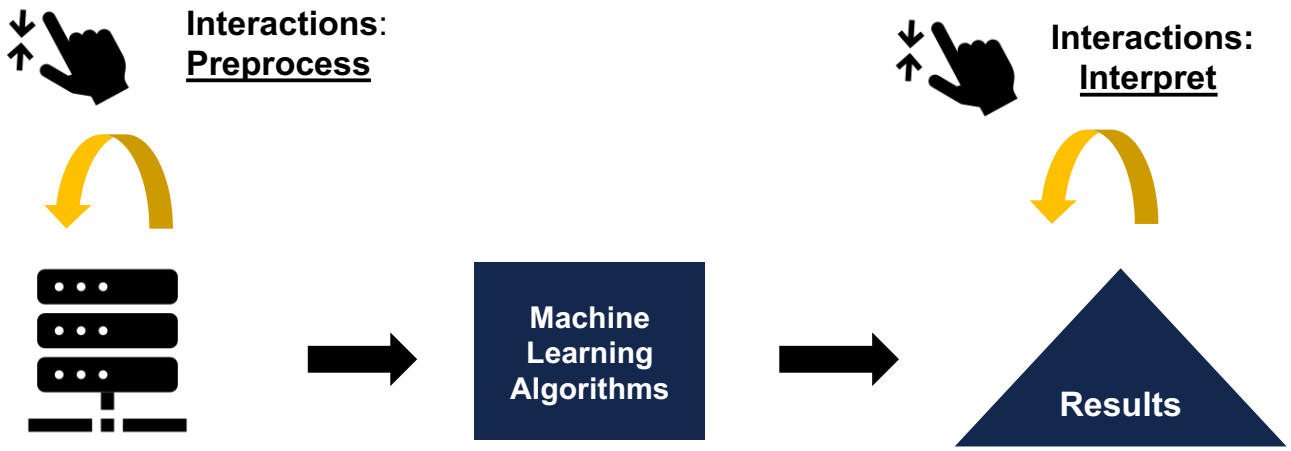
Website: www.banyikun.com/wwwtutorial/

ILLINOIS

Interactions in Machine Learning



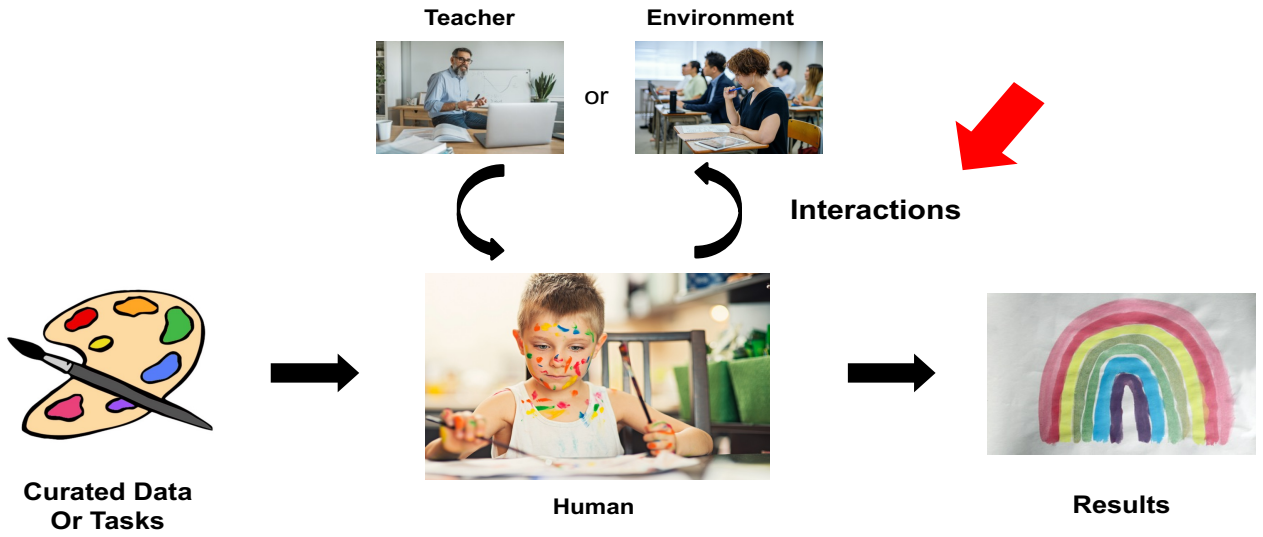
Interactive Learning (Human)



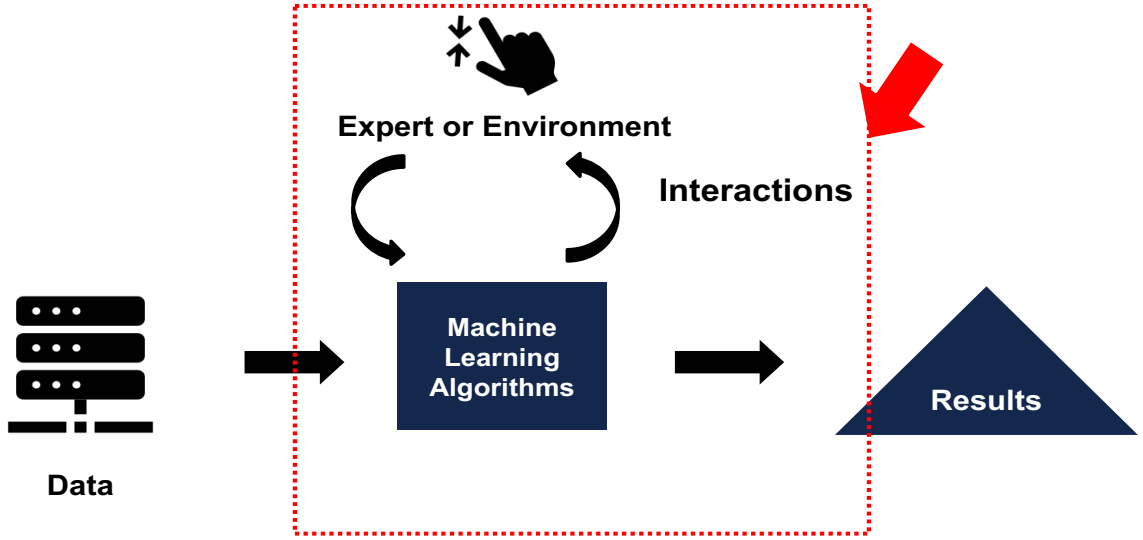
Machine Learning (Conventional)

- Ernst, Damien, and Arthur Louette. "Introduction to reinforcement learning." 2024.
- Fails, Jerry Alan, and Dan R. Olsen Jr. "Interactive machine learning." *Proceedings of the 8th international conference on Intelligent user interfaces*. 2003.
- Teso, Stefano, and Kristian Kersting. "Explanatory interactive machine learning." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

Interactions in Machine Learning



Interactive Learning (Human)



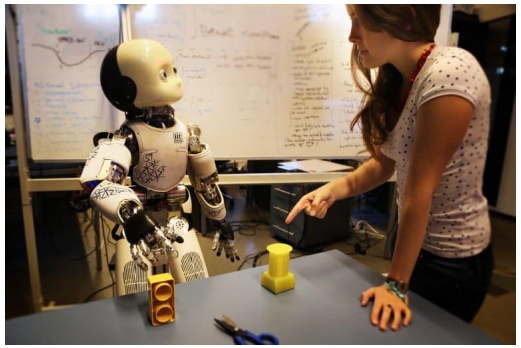
Interactive Machine Learning

- Ernst, Damien, and Arthur Louette. "Introduction to reinforcement learning." 2024.
- Fails, Jerry Alan, and Dan R. Olsen Jr. "Interactive machine learning." *Proceedings of the 8th international conference on Intelligent user interfaces*. 2003.
- Teso, Stefano, and Kristian Kersting. "Explanatory interactive machine learning." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

Interactive Machine Learning and Applications

- **Interactive Machine Learning (IML)** is the core of **Artificial Intelligence (AI)**.

(1) **Recommender Systems**



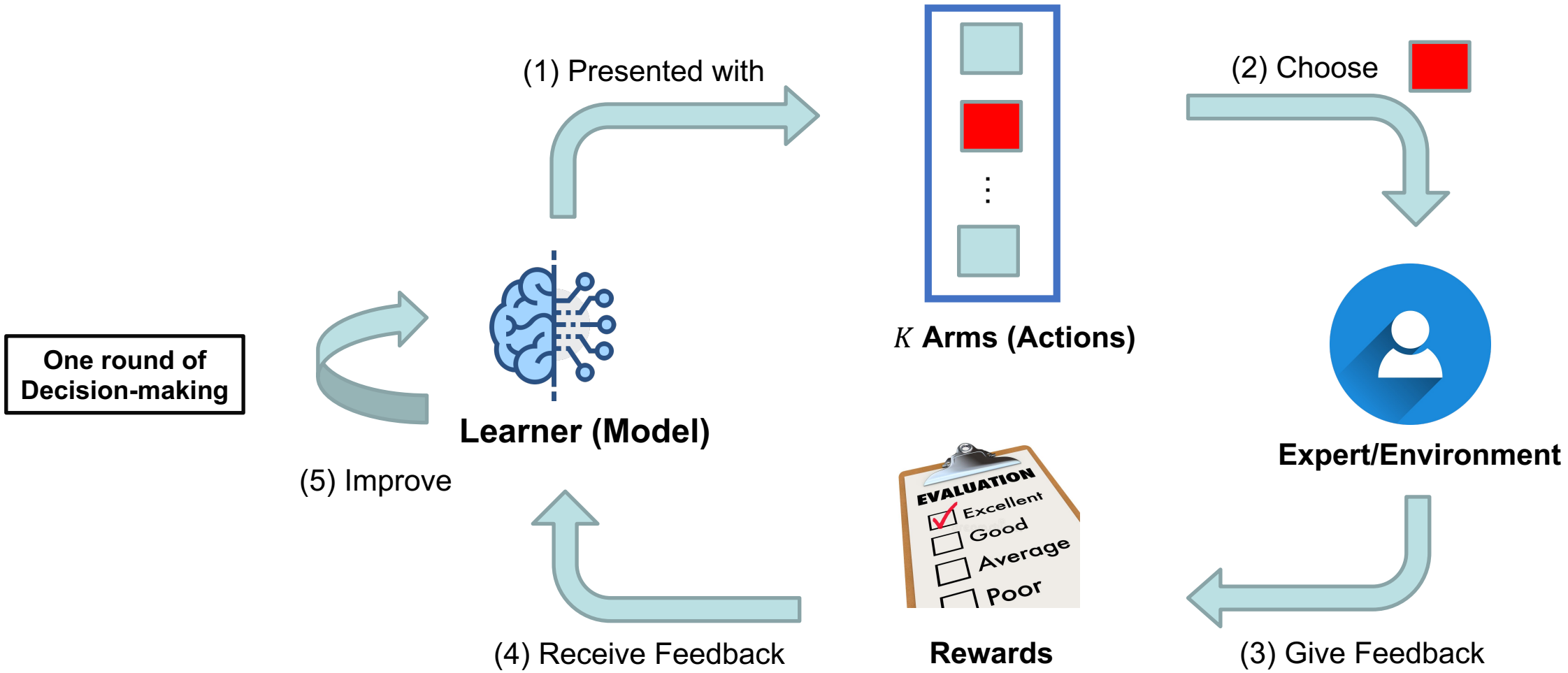
(2) **Robot Learning**

(3) **Language Model**

- Ernst, Damien, and Arthur Louette. "Introduction to reinforcement learning." 2024.
- Fails, Jerry Alan, and Dan R. Olsen Jr. "Interactive machine learning." *Proceedings of the 8th international conference on Intelligent user interfaces*. 2003.
- Teso, Stefano, and Kristian Kersting. "Explanatory interactive machine learning." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

Sequential Decision-Making: Bandits Formulation

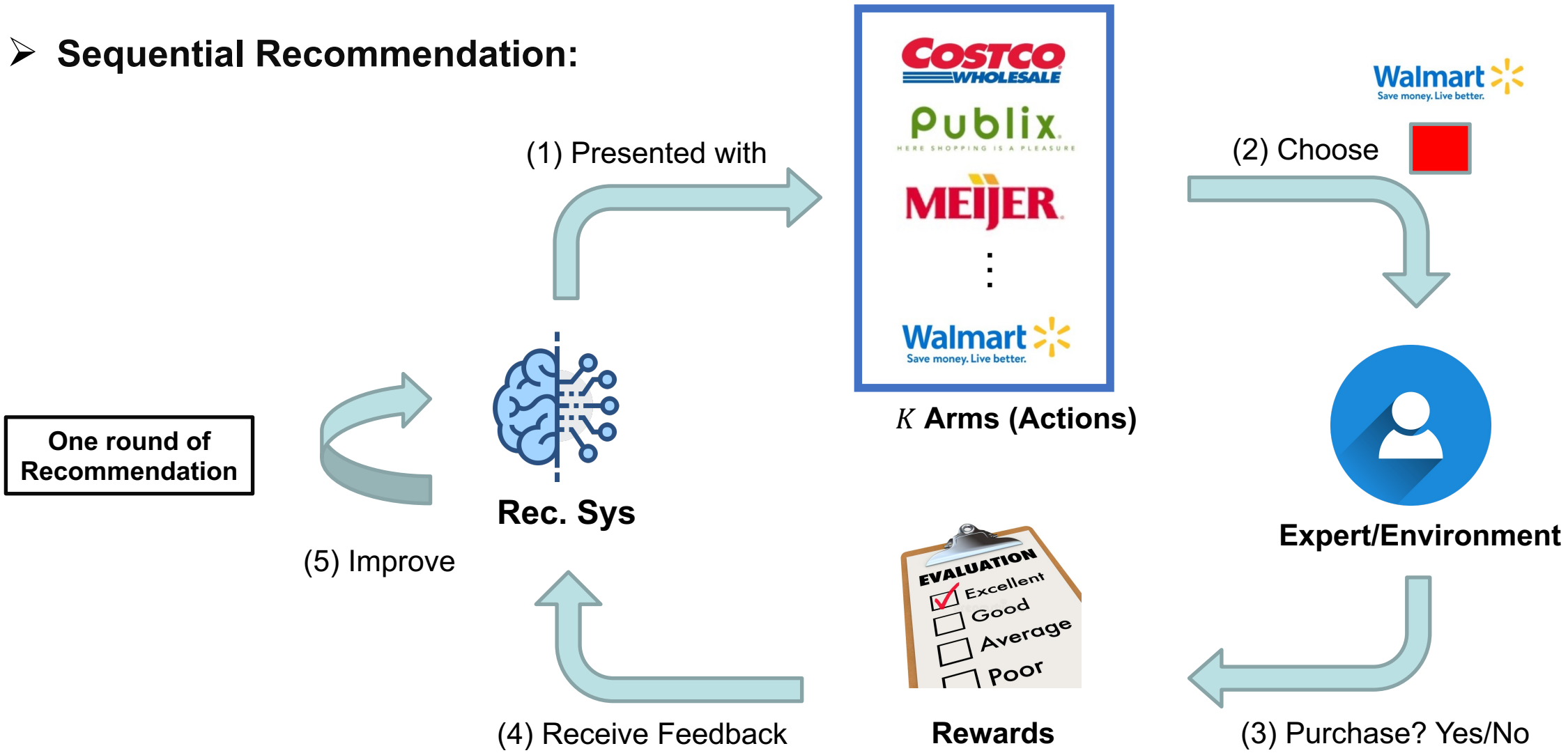
- Many IML scenarios can be formulated as **sequential decision-making**.



- Ernst, Damien, and Arthur Louette. "Introduction to reinforcement learning." 2024.
- Fails, Jerry Alan, and Dan R. Olsen Jr. "Interactive machine learning." *Proceedings of the 8th international conference on Intelligent user interfaces*. 2003.
- Teso, Stefano, and Kristian Kersting. "Explanatory interactive machine learning." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

Sequential Recommendation: Bandits Formulation

➢ **Sequential Recommendation:**

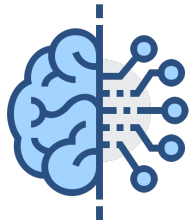


- Ernst, Damien, and Arthur Louette. "Introduction to reinforcement learning." 2024.
- Fails, Jerry Alan, and Dan R. Olsen Jr. "Interactive machine learning." *Proceedings of the 8th international conference on Intelligent user interfaces*. 2003.
- Teso, Stefano, and Kristian Kersting. "Explanatory interactive machine learning." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

Sequential Recommendation: Objective

Goal: Maximize $\sum_{t=1}^{T} \mathbb{E}[r_{t,i_t}]$ **Or** Minimize $\sum_{t=1}^{T} (\mathbb{E}[r_{t,i_t^*}] - \mathbb{E}[r_{t,i_t}])$, where $i_t^* = \arg\max_{i \in [K]} \mathbb{E}[r_{t,i}]$.

Round t:



Rec. Sys

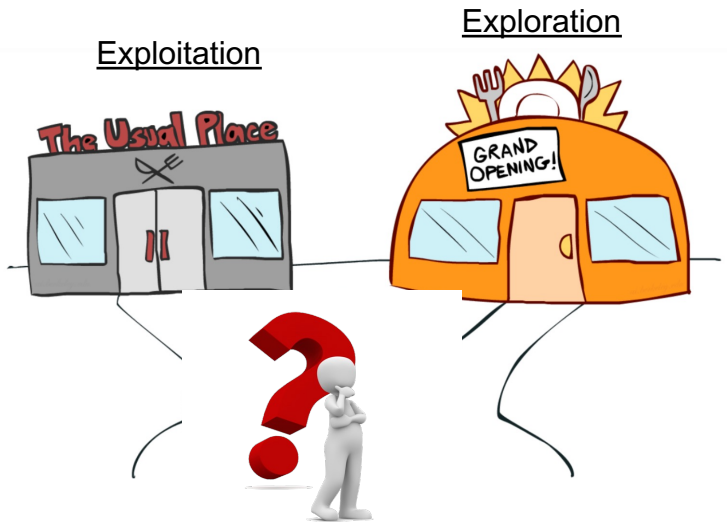
$r_{t,1}$	$r_{t,2}$	$r_{t,3}$	$r_{t,4}$	…	$r_{t,K}$
COSTCO WHOLESALE	Publix. HERE SHOPPING IS A PLEASURE	MEIJER	Walmart Save money. Live better.	…	
Arm 1	Arm 2	Arm 3	Arm 4	…	Arm K

Choose Arm i_t, Observe r_{t,i_t}

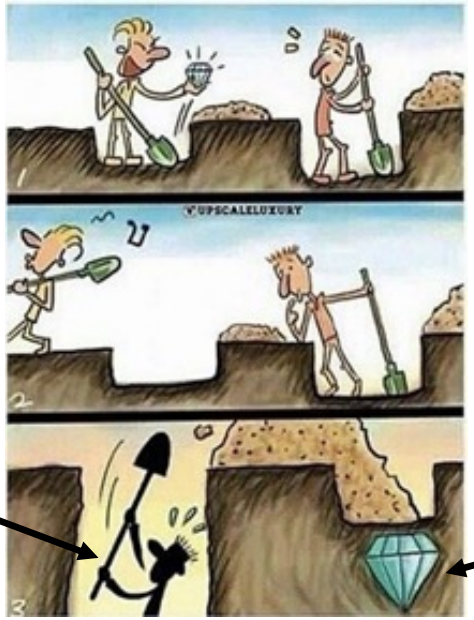
• Ernst, Damien, and Arthur Louette. "Introduction to reinforcement learning." 2024.

Exploitation VS Exploration in Sequential Decision-Making

- Dilemma of **exploitation** and **exploration** is ubiquitous in **human decision-making**.



(1)



(2)

Exploitation:
Exploit past data or observations.
E.g., estimation by a greedy model

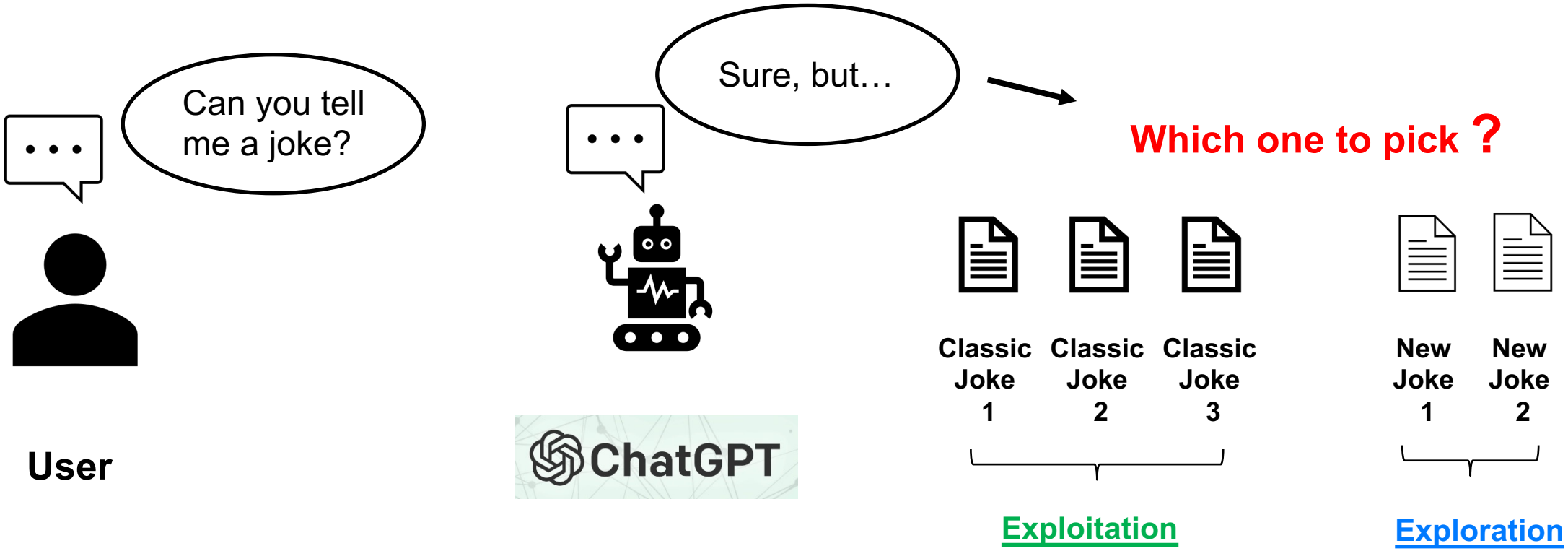
VS

Exploration:
Explore new knowledge for long-term benefit.
E.g., take uncertain actions

• Slivkins, Aleksandrs. "Introduction to multi-armed bandits." *Foundations and Trends® in Machine Learning* 12.1-2 (2019): 1-286.

Exploitation VS Exploration in Sequential Recommendation

- Dilemma of **exploitation** and **exploration** is a **fundamental problem** in **sequential decision-making**.

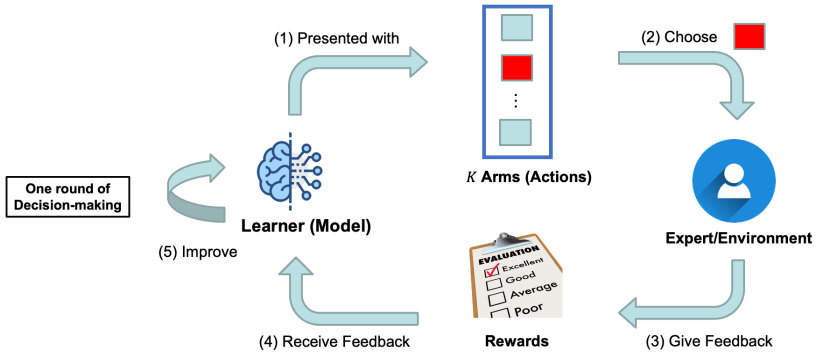


• Slivkins, Aleksandrs. "Introduction to multi-armed bandits." *Foundations and Trends® in Machine Learning* 12.1-2 (2019): 1-286.

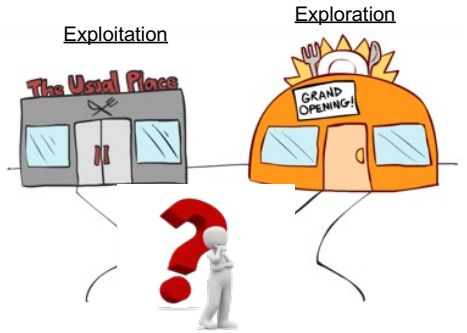
Advantages of Bandit-based Methods

➢ **No Requirement for Large Collected Data**

➢ **Adapt over Time**

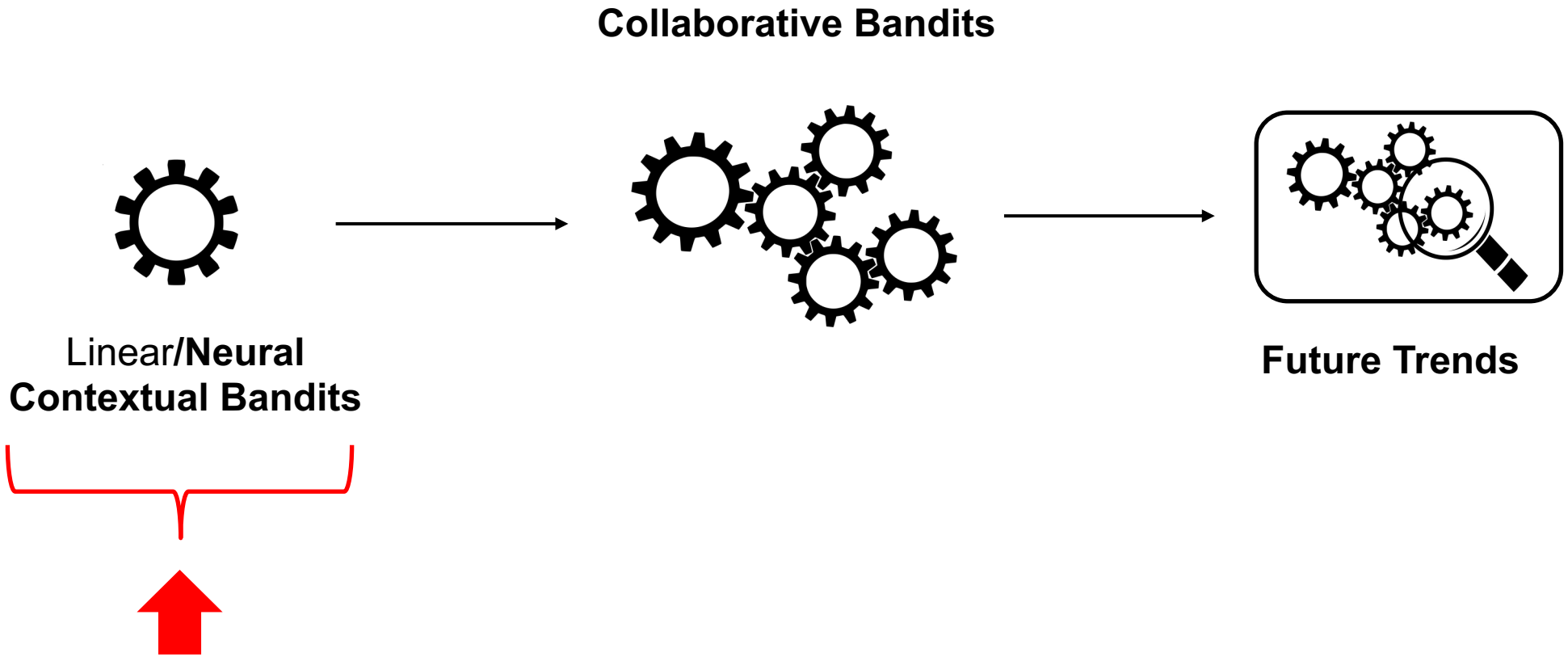


➢ **Explicit Exploration**

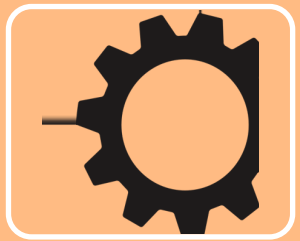


- Ban, Yikun, and Jingrui He. "Local clustering in contextual multi-armed bandits." WWW 2021.
- Gao, Chongming, et al. "Alleviating matthew effect of offline reinforcement learning in interactive recommendation." ICLR 2023.

Roadmap


Collaborative Bandits
Linear/Neural
Contextual Bandits
Future Trends

Neural Contextual Bandits: Roadmap



Fundamental Exploration

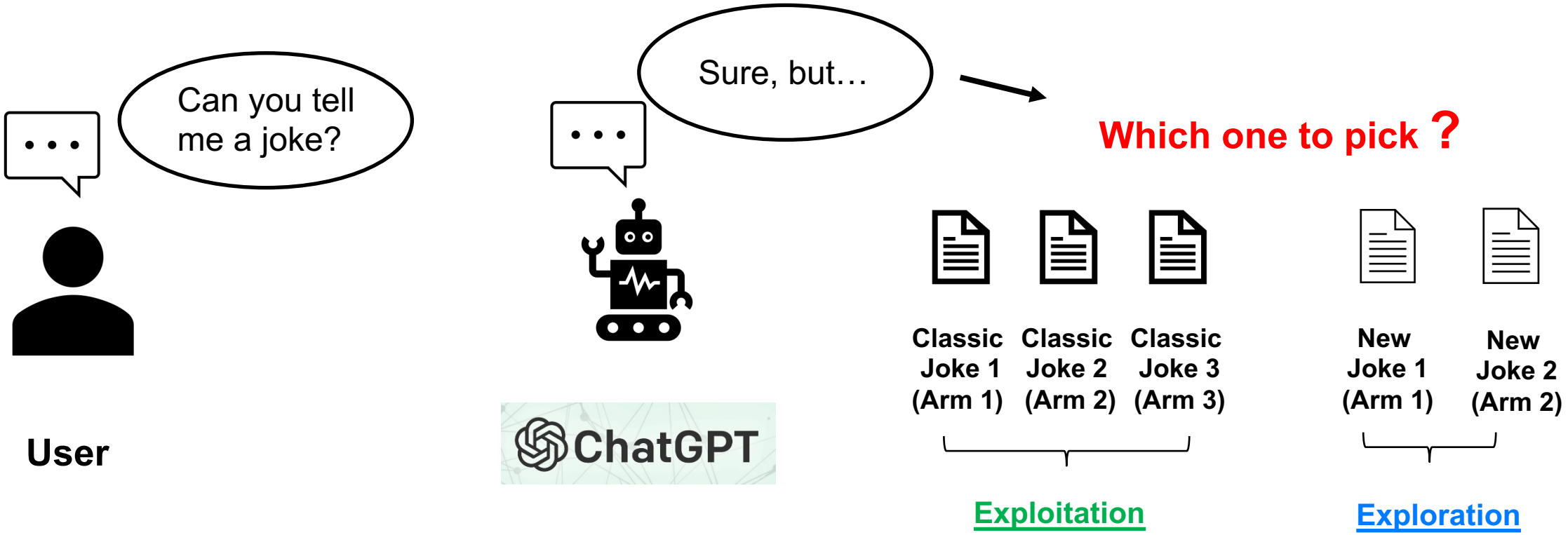
- Upper Confidence Bound
- Thompson Sampling
- Exploration Network

Efficient Exploration

- Neural Linear UCB
- Neural Network with Perturbed Reward
- Inverse Weight Gap Strategy

Background

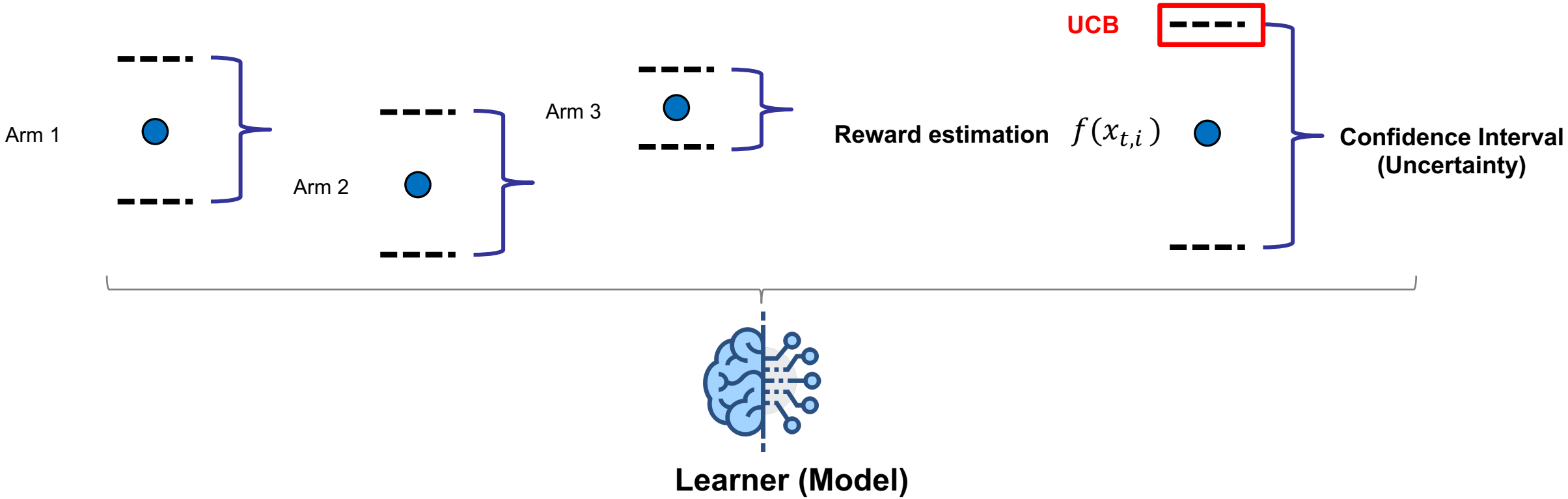
- Popular existing exploration strategies.
  - **ϵ-greedy**: With probability $1-\epsilon$, greedily choose one arm according to history; Otherwise, choose an arm randomly.



• Slivkins, Aleksandrs. "Introduction to multi-armed bandits." *Foundations and Trends® in Machine Learning* 12.1-2 (2019): 1-286.

Background

- Popular existing exploration strategies.
  - **ϵ-greedy**: With probability $1-\epsilon$, greedily choose one arm according to history; Otherwise, choose an arm randomly.
  - **Upper Confidence Bound:**

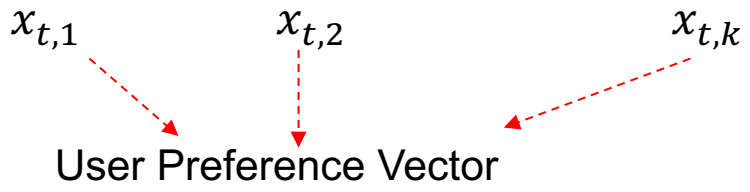


• Slivkins, Aleksandrs. "Introduction to multi-armed bandits." *Foundations and Trends® in Machine Learning* 12.1-2 (2019): 1-286.

Linear UCB: Joint Problem Definition

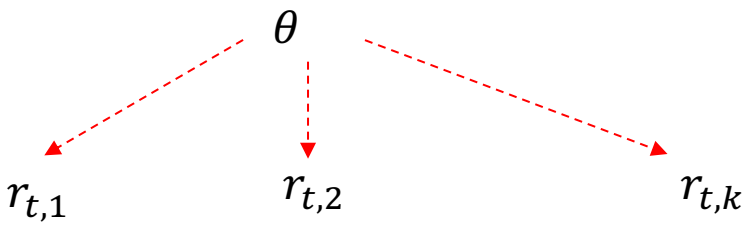
In round t: A user is serving

- **Arm Pool**: Arm 1, Arm 2, ..., Arm k — d-dimensional Feature Vector



- **User Pool**: d-dimensional Preference Vector (**Unknown**)

- **Reward**: Linear reward function



$$r_{t,i} = \theta^T x_{t,i} + \eta_{t,i}$$

Noise, ν-sub-Gaussian

- Chu, Wei, et al. "Contextual bandits with linear payoff functions." AISTATS 2011.
- Li, Lihong, Wei Chu, John Langford, and Robert E. Schapire. "A contextual-bandit approach to personalized news article recommendation." WWW 2010.

Linear UCB: Disjoint Problem Definition

In round t: A user is serving

- **Arm Pool**: Arm 1 (COSTCO WHOLESALE), Arm 2 (Publix), ..., Arm k (Walmart) — d-dimensional Feature Vector

$x_{t,1}$ $x_{t,2}$ $x_{t,k}$

- **User Pool**: User 1, User 2, ..., User n — d-dimensional Preference Vector (**Unknown**)

θ_1 θ_2 θ_n

- **Reward**: $r_{t,1}$ $r_{t,2}$ $r_{t,k}$ — Linear reward function

Given User j,

$$r_{t,i} = \theta_j^T x_{t,i} + \eta_{t,i}$$

Noise, ν-sub-Gaussian

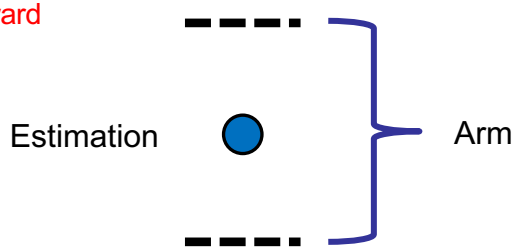
- Chu, Wei, et al. "Contextual bandits with linear payoff functions." AISTATS 2011.
- Li, Lihong, Wei Chu, John Langford, and Robert E. Schapire. "A contextual-bandit approach to personalized news article recommendation." WWW 2010.

Linear UCB

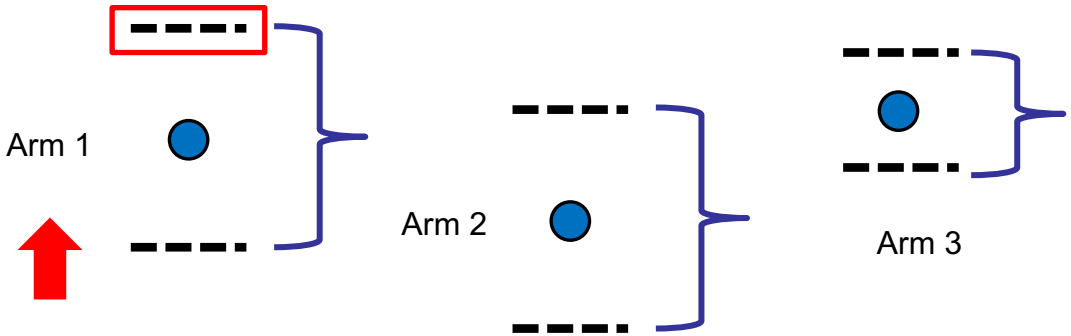
Confidence Interval:

$$\left|\hat{r}_{t,a} - x_{t,a}^{\top}\theta^*\right| \leq (\alpha+1)s_{t,a}.$$

Estimated Expected Reward



$$a_t \overset{\text{def}}{=} \arg\max_{a \in \mathcal{A}_t} \left(\mathbf{x}_{t,a}^{\top}\hat{\boldsymbol{\theta}}_a + \alpha\sqrt{\mathbf{x}_{t,a}^{\top}\mathbf{A}_a^{-1}\mathbf{x}_{t,a}} \right),$$



Joint Linear Models

Estimated by Ridge Regression

for $t = 1, 2, 3, \ldots, T$ **do** -- A user is serving in each round

$\theta_t \leftarrow A^{-1}b$

(Item 1) (Item 2) ⋯

Observe K features, $x_{t,1}, x_{t,2}, \cdots, x_{t,K} \in \mathbb{R}^d$

for $a = 1, 2, \ldots, K$ **do**

$p_{t,a} \leftarrow \theta_t^{\top} x_{t,a} + \alpha\sqrt{x_{t,a}^{\top} A^{-1} x_{t,a}}$ {Computes upper confidence bound}

end for

Exploitation

Exploration

Choose action $a_t = \arg\max_a p_{t,a}$ with ties broken arbitrarily

Observe payoff $r_t \in \{0, 1\}$

$A \leftarrow A + x_{t,a_t} x_{t,a_t}^{\top}$

$b \leftarrow b + x_{t,a_t} r_t$

Estimated by Ridge Regression

end for

- Chu, Wei, et al. "Contextual bandits with linear payoff functions." AISTATS 2011.
- Li, Lihong, Wei Chu, John Langford, and Robert E. Schapire. "A contextual-bandit approach to personalized news article recommendation." WWW 2010.

Linear UCB: Regret Analysis

➢ Confidence Interval:

Estimated by Ridge Regression

With high probability,

$$\left|\hat{r}_{t,a} - x_{t,a}^{\top}\theta^*\right| \leq (\alpha+1)s_{t,a}.$$

where

$$s_{t,a} = \sqrt{x_{t,a}^{\top}A_t^{-1}x_{t,a}} \in \mathbb{R}_+$$

$$\begin{aligned}
&\hat{r}_{t,a} - x_{t,a}^{\top}\theta^* \\
&= x_{t,a}^{\top}\theta_t - x_{t,a}^{\top}\theta^* \\
&= x_{t,a}^{\top}A_t^{-1}b_t - x_{t,a}^{\top}A_t^{-1}\left(I_d + D_t^{\top}D_t\right)\theta^* \\
&= x_{t,a}^{\top}A_t^{-1}D_t^{\top}y_t - x_{t,a}^{\top}A_t^{-1}\left(\theta^* + D_t^{\top}D_t\theta^*\right) \\
&= x_{t,a}^{\top}A_t^{-1}D_t^{\top}\left(y_t - D_t\theta^*\right) - x_{t,a}^{\top}A_t^{-1}\theta^*,
\end{aligned}$$

➢ Regret Upper Bound

$$O\left(\sqrt{Td\ln^3(KT\ln(T)/\delta)}\right).$$

The Number of Rounds

Dimensionality of Item Context Vector

The Number of Items

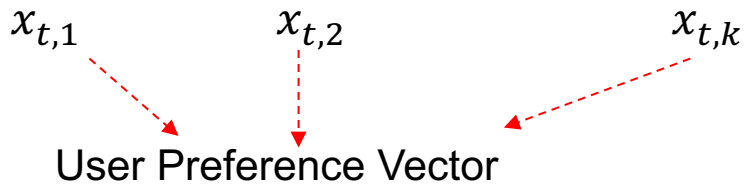
- Chu, Wei, et al. "Contextual bandits with linear payoff functions." AISTATS 2011.
- Li, Lihong, Wei Chu, John Langford, and Robert E. Schapire. "A contextual-bandit approach to personalized news article recommendation." WWW 2010.

Neural Bandits: Problem Formulation

In round t:

- **Arm Pool**:

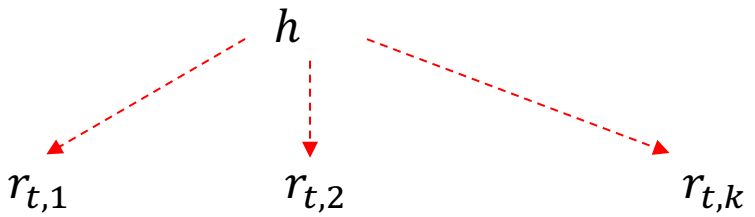
d-dimensional Feature Vector



- **User Pool**:

Preference Function (**Unknown**)

- **Reward**:



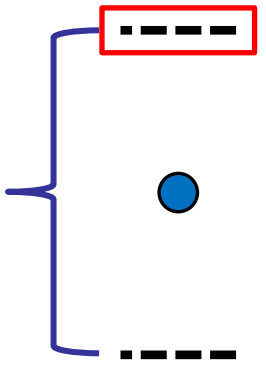
General reward function (Linear/Non-Linear)

$$r_{t,i} = h(x_{t,i}) + \eta_{t,i}$$

Sub-Gaussian Noise

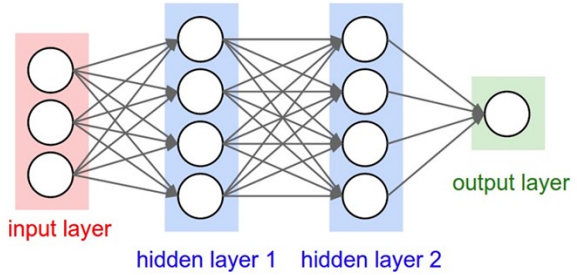
• Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML 2020.

Neural UCB: Method



$$U_{t,a}$$

$$f(\mathbf{x};\boldsymbol{\theta}) = \sqrt{m}\mathbf{W}_L\sigma\Big(\mathbf{W}_{L-1}\sigma\big(\cdots\sigma(\mathbf{W}_1\mathbf{x})\big)\Big)$$



Gradient of f

$$U_{t,a} = \underbrace{f(\mathbf{x}_{t,a};\boldsymbol{\theta}_{t-1})}_{\text{mean}} + \gamma_{t-1}\underbrace{\sqrt{\mathbf{g}(\mathbf{x}_{t,a};\boldsymbol{\theta}_{t-1})^\top\mathbf{Z}_{t-1}^{-1}\mathbf{g}(\mathbf{x}_{t,a};\boldsymbol{\theta}_{t-1})/m}}_{\text{variance}}$$

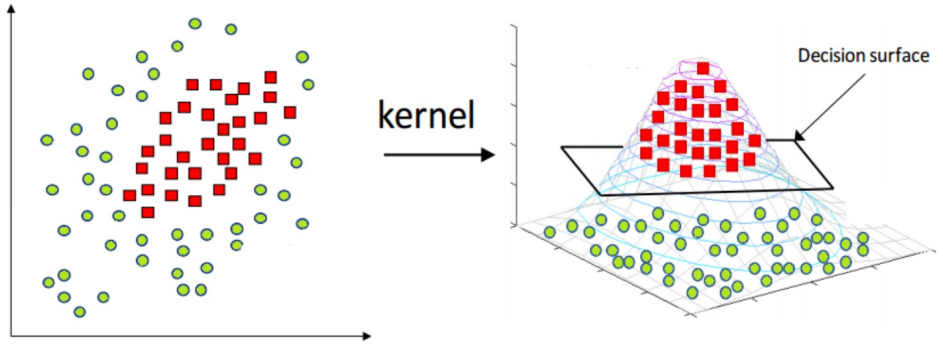
Compared with LinUCB (Li et al. 2010)

$$U_{t,a} = \underbrace{\langle\mathbf{x}_{t,a},\boldsymbol{\theta}_{t-1}\rangle}_{\text{mean}} + \gamma_{t-1}\underbrace{\sqrt{\mathbf{x}_{t,a}^\top\mathbf{Z}_{t-1}^{-1}\mathbf{x}_{t,a}}}_{\text{variance}}$$

• Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML 2020.

Neural Tangent Kernel

- A **sufficiently wide neural network** behaves like a **linearized model** governed by the derivative of network with respect to its parameters (**Gradient**).

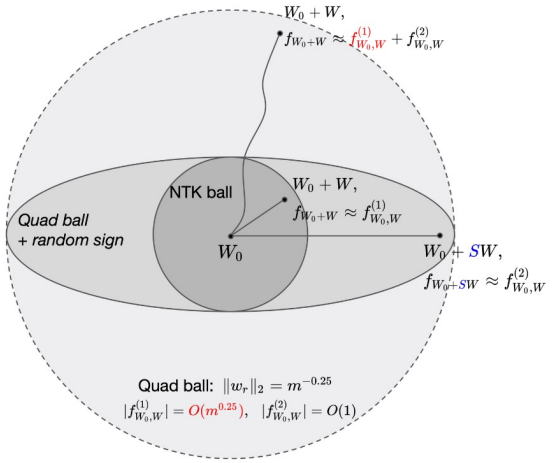


https://www.geeksforgeeks.org/major-kernel-functions-in-support-vector-machine-svm/

Neural Tangent Kernel

$$\Theta(x, x'; \theta) = \nabla_\theta f(x; \theta) \cdot \nabla_\theta f(x'; \theta).$$

- With near-**infinite width**, Neural network behave like a **kernel predictor** with Neural Tangent Kernel (NTK)



- Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." NeurIPS 2018.
- Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." ICML 2019.

Neural UCB: Workflow

- In each round, a user is serving

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

K arms (Items)

Observe $\{\mathbf{x}_{t,a}\}_{a=1}^{K}$

for $a = 1, \ldots, K$ **do**

Exploitation

Exploration

Compute $U_{t,a} = f(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\sqrt{\mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})^\top \mathbf{Z}_{t-1}^{-1} \mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})/m}$

Let $a_t = \text{argmax}_{a \in [K]} U_{t,a}$

end for

Play a_t and observe reward r_{t,a_t}

Compute $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})\mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})^\top / m$

Similar to Linear Regression

Let $\boldsymbol{\theta}_t = \text{TrainNN}(\lambda, \eta, J, m, \{\mathbf{x}_{i,a_i}\}_{i=1}^{t}, \{r_{i,a_i}\}_{i=1}^{t}, \boldsymbol{\theta}_0)$

Train Neural Networks

Compute

$$\gamma_t = \sqrt{1 + C_1 m^{-1/6}\sqrt{\log m} L^4 t^{7/6} \lambda^{-7/6}} \cdot \left(\nu\sqrt{\log \frac{\det \mathbf{Z}_t}{\det \lambda \mathbf{I}} + C_2 m^{-1/6}\sqrt{\log m} L^4 t^{5/3} \lambda^{-1/6} - 2\log\delta} + \sqrt{\lambda} S\right)$$

$$+ (\lambda + C_3 t L)\left[(1 - \eta m \lambda)^{J/2}\sqrt{t/\lambda} + m^{-1/6}\sqrt{\log m} L^{7/2} t^{5/3} \lambda^{-5/3}(1 + \sqrt{t/\lambda})\right].$$

end for

Confidence Radius

Neural Function Approximation Error

• Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML 2020.

Neural UCB: Regret Analysis

- Definition of **NTK Matrix** on all observed contexts of T rounds.

$$\widetilde{\mathbf{H}}_{i,j}^{(1)} = \mathbf{\Sigma}_{i,j}^{(1)} = \langle \mathbf{x}^i, \mathbf{x}^j \rangle, \qquad \mathbf{A}_{i,j}^{(l)} = \begin{pmatrix} \mathbf{\Sigma}_{i,i}^{(l)} & \mathbf{\Sigma}_{i,j}^{(l)} \\ \mathbf{\Sigma}_{i,j}^{(l)} & \mathbf{\Sigma}_{j,j}^{(l)} \end{pmatrix},$$

$$\mathbf{\Sigma}_{i,j}^{(l+1)} = 2\mathbb{E}_{(u,v)\sim N(\mathbf{0},\mathbf{A}_{i,j}^{(l)})}\left[\sigma(u)\sigma(v)\right],$$

$$\widetilde{\mathbf{H}}_{i,j}^{(l+1)} = 2\widetilde{\mathbf{H}}_{i,j}^{(l)}\mathbb{E}_{(u,v)\sim N(\mathbf{0},\mathbf{A}_{i,j}^{(l)})}\left[\sigma'(u)\sigma'(v)\right] + \mathbf{\Sigma}_{i,j}^{(l+1)}.$$

Then, $\mathbf{H} = (\widetilde{\mathbf{H}}^{(L)} + \mathbf{\Sigma}^{(L)})/2$ is called the *neural tangent kernel (NTK)* matrix on the context set.

- Analyze dynamics of gradient and NTK regression.

Assumption: $\mathbf{H} \succeq \lambda_0 \mathbf{I}$.

- Satisfied if no two observed arm contexts are parallel.

Lemma: When neural network is wide enough,

$$h(\mathbf{x}^i) = \langle \mathbf{g}(\mathbf{x}^i; \boldsymbol{\theta}_0), \boldsymbol{\theta}^* - \boldsymbol{\theta}_0 \rangle,$$

Linear function w.r.t. Gradient

$$\sqrt{m}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 \leq \sqrt{2\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}, \qquad (5.1)$$

for all $i \in [TK]$.

- Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." NeurIPS 2018.
- Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML 2020.

Neural UCB: Regret Analysis

Assumption: $\mathbf{H} \succeq \lambda_0 \mathbf{I}$.

Satisfied if *no* two contexts in $\{\mathbf{x}^i\}_{i=1}^{TK}$ are parallel.

$$h(\mathbf{x}^i) = \langle \mathbf{g}(\mathbf{x}^i; \boldsymbol{\theta}_0), \boldsymbol{\theta}^* - \boldsymbol{\theta}_0 \rangle,$$

$$\sqrt{m}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 \leq \sqrt{2\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}},$$

$$S = \sqrt{2\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}$$

$$\tilde{d} = \frac{\log\det(\mathbf{I} + \mathbf{H}/\lambda)}{\log(1 + TK/\lambda)}$$

LinUCB:

$\widetilde{O}(d\sqrt{T})$

Theorem

Let $\mathbf{h} = [h(\mathbf{x}^i)]_{i=1}^{TK} \in \mathbb{R}^{TK}$. *Set* $J = \widetilde{\boldsymbol{\Theta}}(TL/\lambda)$, $\eta = \Theta((mTL + m\lambda)^{-1})$ *and* $S = 2\sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}$. *Under the overparameterized setting* ($m \gg 1$), *with probability at least* $1 - \delta$,

$$R_T = \widetilde{O}\Big(\sqrt{\tilde{d}T}\sqrt{\max\{\tilde{d}, S^2\}}\Big).$$

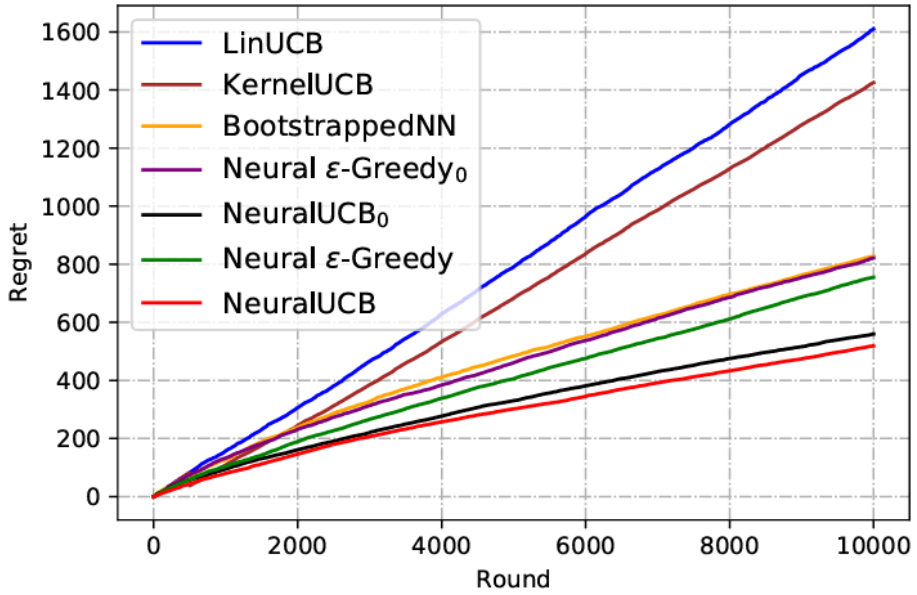
Upper Bound of Neural Parameters

Effective dimension in NTK Space

• Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML 2020.

Neural UCB: Empirical Evaluation

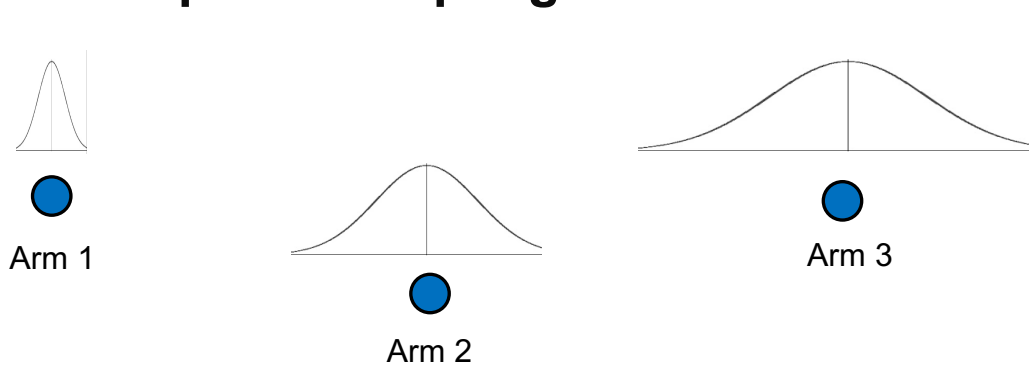
- NeuralUCB uses neural networks for exploitation, and gradient to explore.
- NeuralUCB achieve $\tilde{O}(\sqrt{T})$ regret upper bound, similar to LinearUCB.
- NeuralUCB generally outperforms linear contextual bandits.

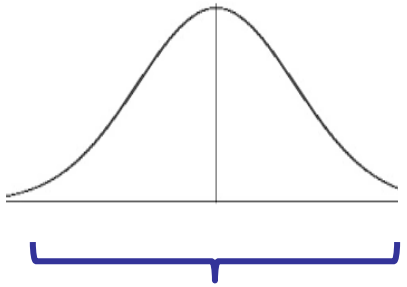


• Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML 2020.

Thompson Sampling

- Popular existing exploration strategies.
  - **ϵ-greedy**: With probability $1-\epsilon$, greedily choose one arm according to history; Otherwise, choose an arm randomly.
  - **Upper Confidence Bound[1].**
  - **Thompson Sampling[2]:**



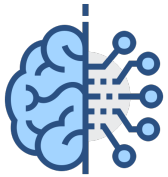


$f(x_{t,i})$ Reward Estimation

ν Standard Deviation (Uncertainty)

Draw from $D(f, \nu)$

E.g., normal distribution

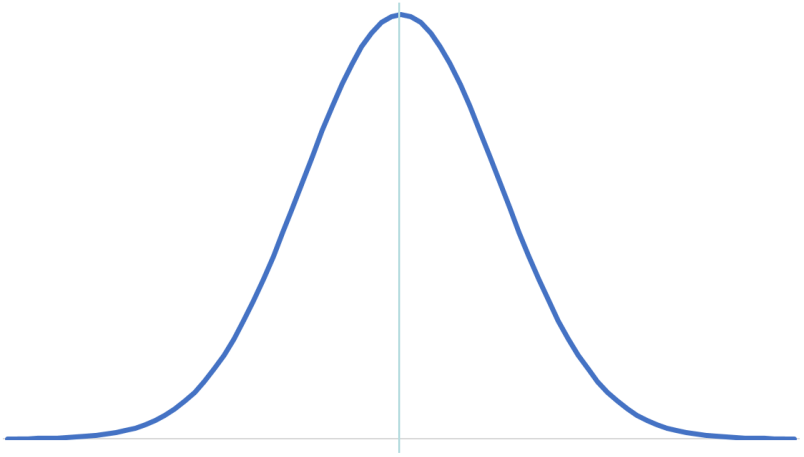


Learner (Model)

[1] Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML, 2020.
[2] Zhang, Weitong, et al. "Neural thompson sampling." ICLR 2021.

Linear Thompson Sampling

Reward Distribution (Gaussian Prior)



User Preference Parameter (Unknown)

$$\mathcal{N}(b_i(t)^T \mu, v^2)$$

Arm context

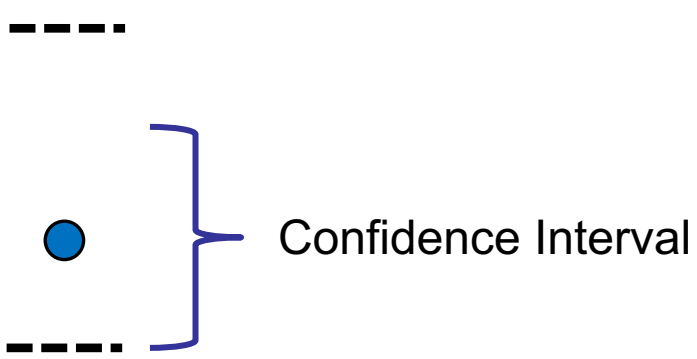
Arm

for all $t = 1, 2, \ldots,$ **do**
  Sample $\tilde{\mu}(t)$ from distribution $\mathcal{N}(\hat{\mu}, v^2 B^{-1})$.
  Play arm $a(t) := \arg\max_i b_i(t)^T \tilde{\mu}(t)$, and observe reward r_t.
  Update $B = B + b_{a(t)}(t) b_{a(t)}(t)^T$, $f = f + b_{a(t)}(t) r_t$, $\hat{\mu} = B^{-1} f$.
end for

Estimated User Preference

Sampled Reward

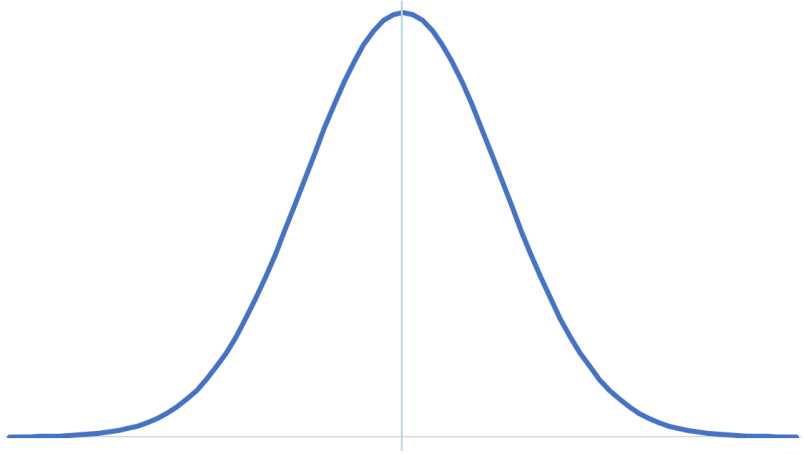
Estimated By Ridge Regression



• Agrawal, Shipra, and Navin Goyal. "Thompson sampling for contextual bandits with linear payoffs." ICML 2013.

Neural Thompson Sampling

Reward Distribution (Gaussian Prior)



$$N(h(x_{t,k}), \nu^2)$$

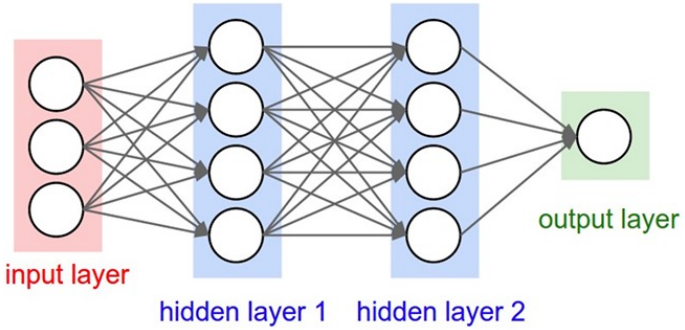
Expected Reward and Variance

Arm

Estimated Distribution:

$$\mathcal{N}(f(\mathbf{x}_{t,k};\boldsymbol{\theta}_{t-1}), \nu^2\sigma_{t,k}^2)$$

$$f(\mathbf{x};\boldsymbol{\theta}) = \sqrt{m}\mathbf{W}_L\sigma\Big(\mathbf{W}_{L-1}\sigma\big(\cdots\sigma(\mathbf{W}_1\mathbf{x})\big)\Big)$$



- Zhang, Weitong, Dongruo Zhou, Lihong Li, and Quanquan Gu. "Neural thompson sampling." ICLR 2021.

Neural Thompson Sampling

- In each round, a user is serving

for $t = 1, \cdots, T$ **do**
  for $k = 1, \cdots, K$ **do**
    $\sigma_{t,k}^2 = \lambda \mathbf{g}^\top(\mathbf{x}_{t,k}; \boldsymbol{\theta}_{t-1}) \mathbf{U}_{t-1}^{-1} \mathbf{g}(\mathbf{x}_{t,k}; \boldsymbol{\theta}_{t-1})/m$
    Sample estimated reward $\widetilde{r}_{t,k} \sim \mathcal{N}(f(\mathbf{x}_{t,k}; \boldsymbol{\theta}_{t-1}), \nu^2 \sigma_{t,k}^2)$
  end for
  Pull arm a_t and receive reward r_{t,a_t}, where $a_t = \text{argmax}_a \widetilde{r}_{t,a}$
  Set $\boldsymbol{\theta}_t$ to be the output of gradient descent for solving (2.3)
  $\mathbf{U}_t = \mathbf{U}_{t-1} + \mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_t)\mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_t)^\top/m$
end for

K arms

Similar to Linear Regression

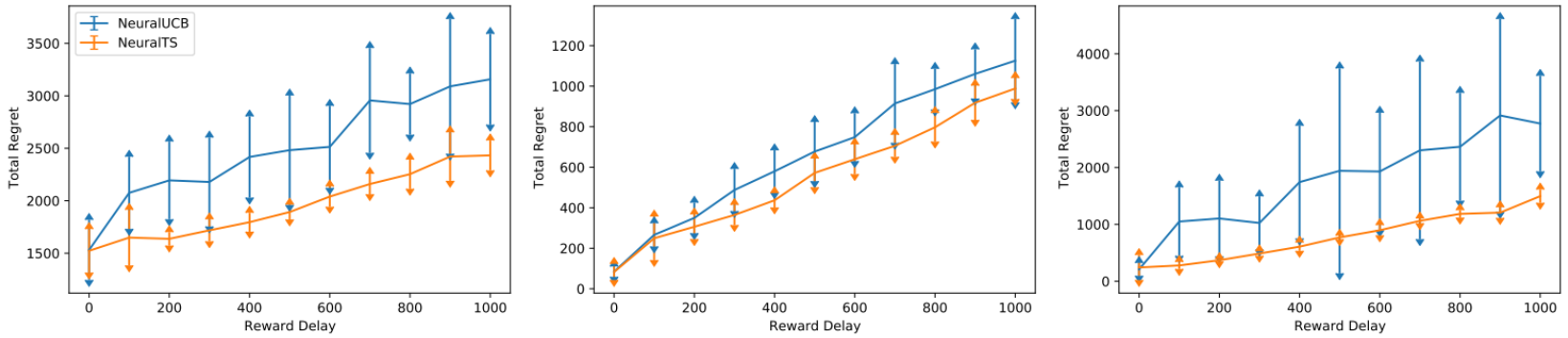
Mean

Variance

Compared to NeuralUCB: $U_{t,a} = f(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\sqrt{\mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})^\top \mathbf{Z}_{t-1}^{-1} \mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})/m}$

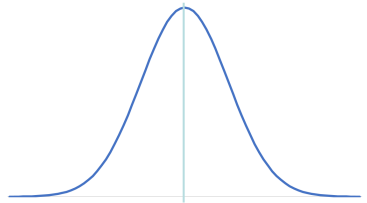
• Zhang, Weitong, Dongruo Zhou, Lihong Li, and Quanquan Gu. "Neural thompson sampling." ICLR 2021.

Neural Thompson Sampling

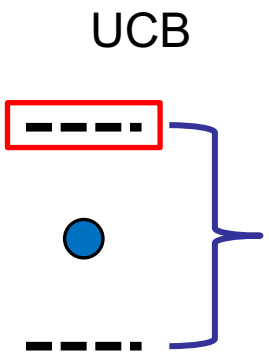


(a) MNIST (b) Mushroom (c) Shuttle

- NerualTS and NeuralUCB have **similar performance** when network is trained every iteration.
- NeuralTS is more robust than NeuralUCB when network is trained **in batch**.
- NeuralTS introduces more **robustness** in exploration.



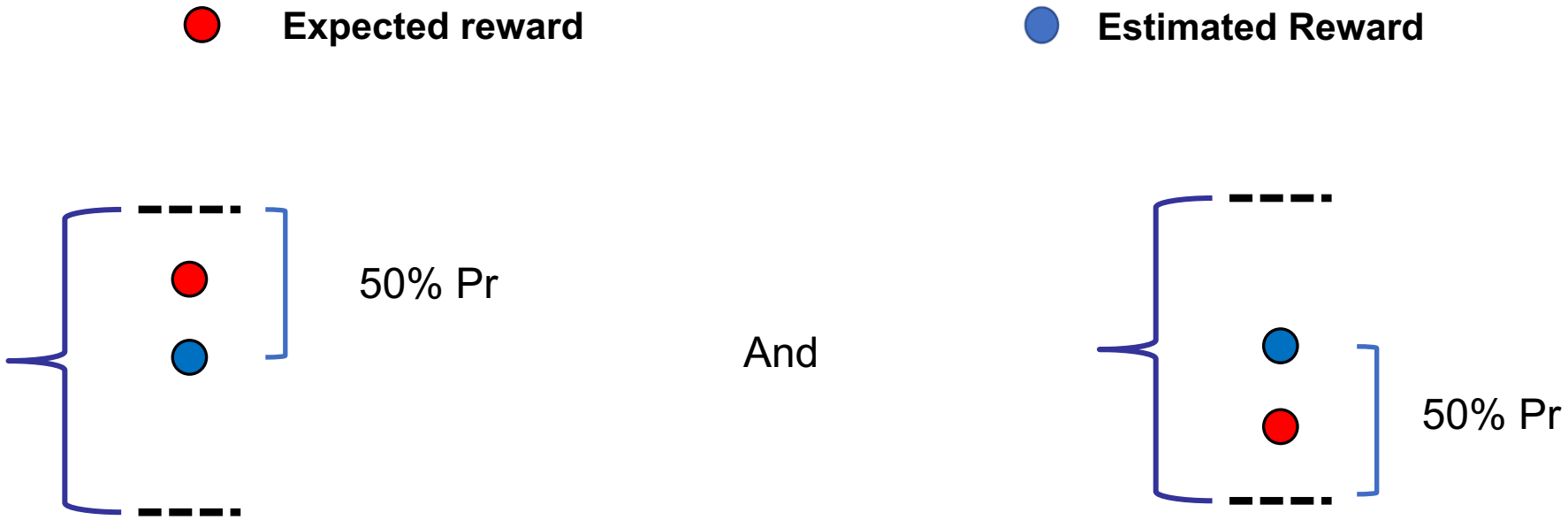
TS



• Zhang, Weitong, Dongruo Zhou, Lihong Li, and Quanquan Gu. "Neural thompson sampling." ICLR 2021.

EE-Net: Background

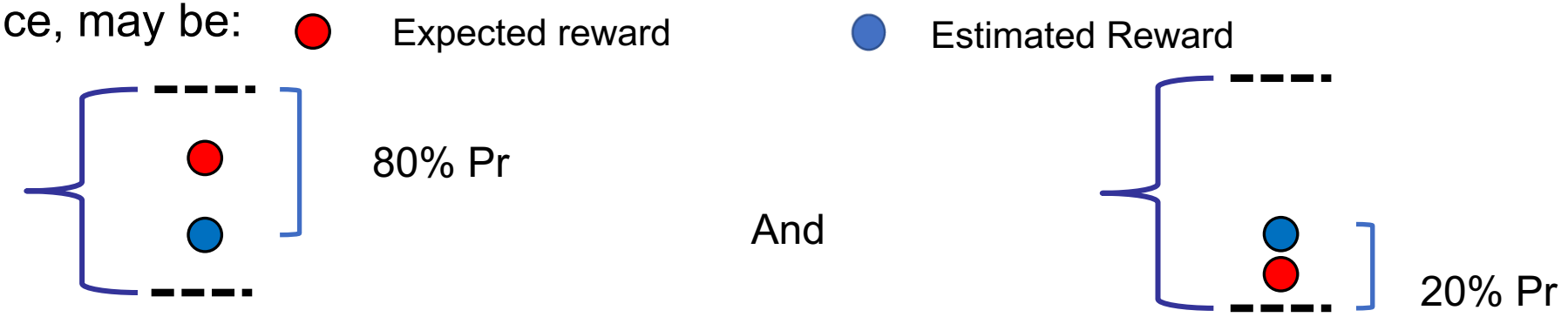
- UCB-based and TS-based exploration highly rely on large-deviation-based **statistical confidence interval.**
  - **Ideal** scenario:



- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Motivation

- UCB-based and TS-based exploration highly rely on large-deviation-based **statistical confidence bound.**
  - In practice, may be:

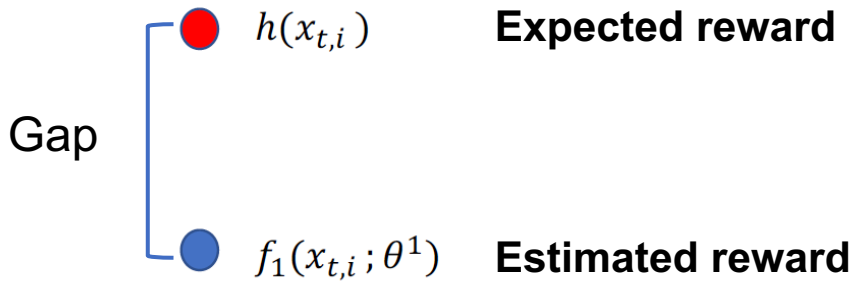


Asymmetric

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Motivation

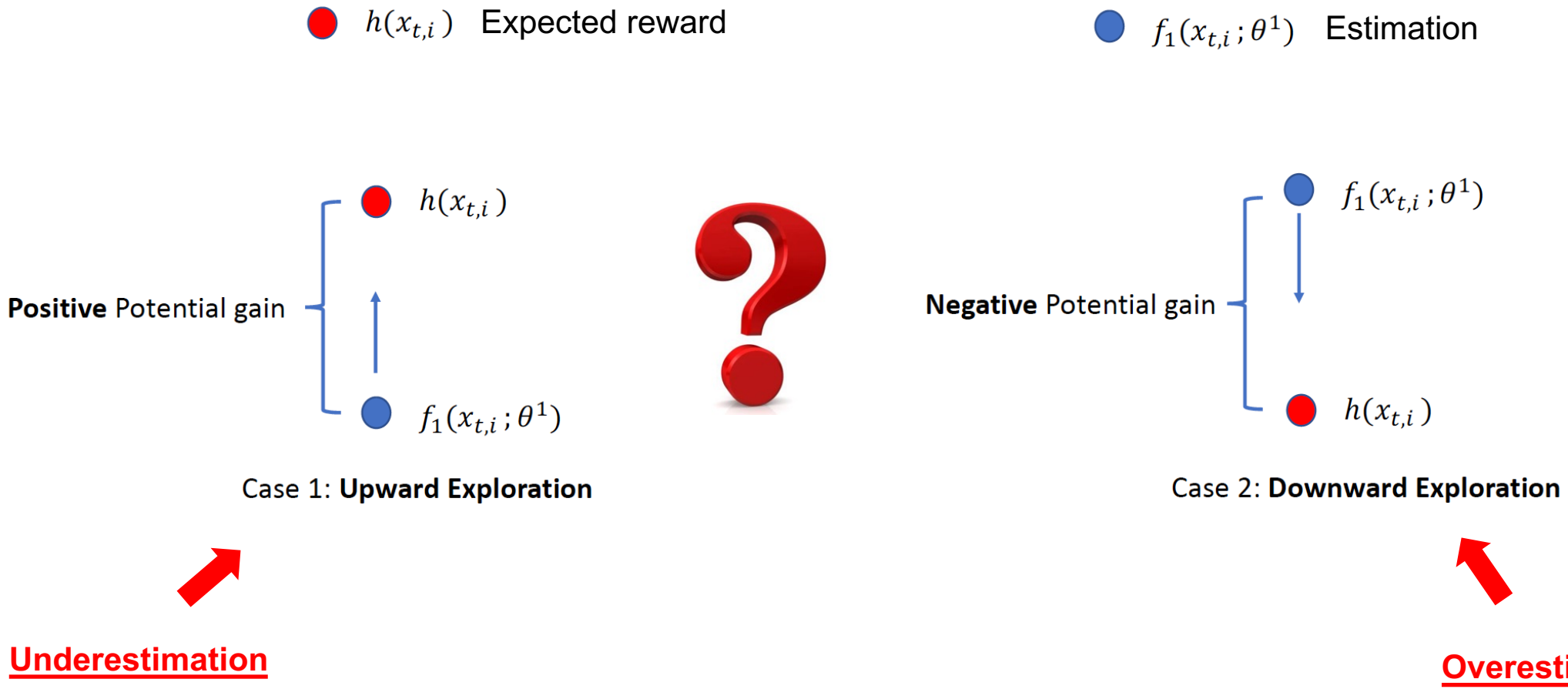
- **Why making exploration?**
  - Because we cannot make accurate prediction on a subset of data.
- **Goal of exploration**: Fill the **gap** between **expected reward** and **estimated reward**.



- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

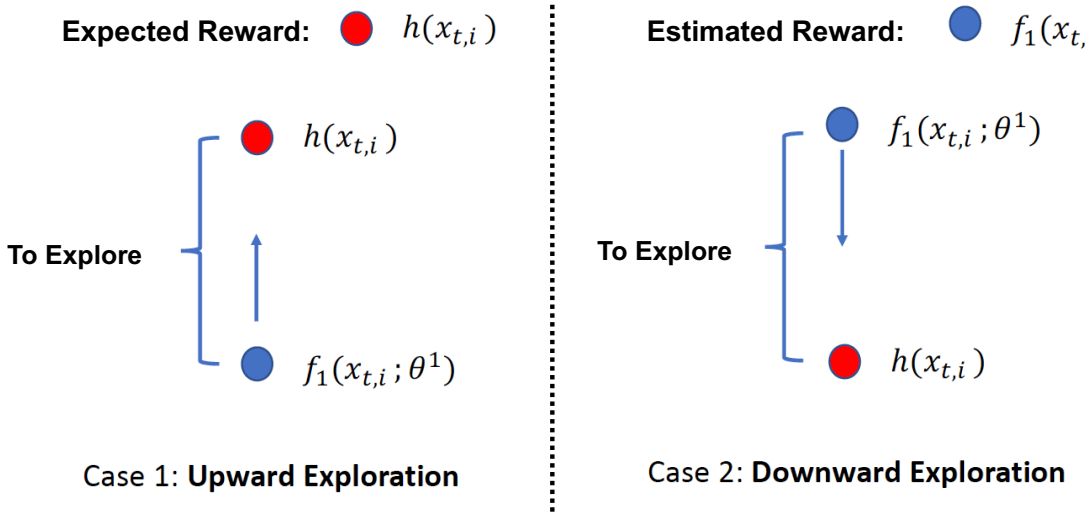
EE-Net: Exploration Direction

➢ **Two types of exploration:** "**Upward**" exploration and "**downward**" Exploration.



- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

Adapt to Exploration Direction is Challenging



Challenge:

Not 50% vs 50% !

Datasets	Upward Exploration	Downward Exploration
Mnist	**76.3%**	23.7%
Disin	29.1%	**70.9%**
MovieLens	**58.6%**	41.4%
Yelp	**55.3%**	44.7%

Pessimistic Model (Human)

Optimistic Model (Human)

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Solution

- Motivation: Can we have an **adaptive** exploration strategy for both “upward” and “downward” exploration?

- **Proposed solution**: We propose to use **another neural network to learn** the gap between expected reward and estimated reward (**potential gain**) **incorporating exploration direction**.

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Motivation and Solution

- Motivation: Can we have an **adaptive** exploration strategy for both "upward" and "downward" exploration?

- **Proposed solution**: We propose to use **another neural network to learn** the gap between expected reward and estimated reward (**potential gain**) **incorporating exploration direction**.

- **Exploitation neural network** f_1 to estimate reward:
  - Given an arm $x_{t,i}$,

$$f_1\left(x_{t,i};\theta^1\right) = \boldsymbol{W}_L\sigma(\boldsymbol{W}_{L-1}\sigma\left(\dots\sigma(\boldsymbol{W}_1\,\cdot)\right))$$

  - $f_1(x_{t,i};\theta^1)$ is to **estimate expected reward** represented by some unknown function $h(x_{t,i})$.
  - In round t, θ^1 is **trained on data of past $t-1$ rounds**, using **gradient descent**.

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Exploration Neural Networks

- **Exploration neural network f_2** (novel component) to estimate potential gain:
  - Given an arm $x_{t,i}$ and its estimation $f_1(x_{t,i};\theta^1)$, **expected potential gain** is defined as:

$$h(x_{t,i}) - f_1(x_{t,i};\theta^1),$$

    where $h(x_{t,i})$ is the expected reward.
  - Thus, given the received reward $r_{t,i}$, **potential gain** is defined as:

$$r_{t,i} - f_1(x_{t,i};\theta^1),$$

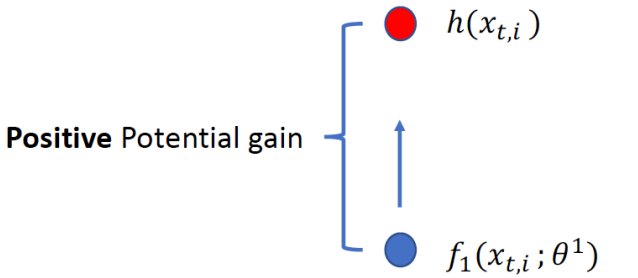
    where $\mathbb{E}[r_{t,i}] = h(x_{t,i})$.
  - Potential gain has a good property: **Indicating exploration direction.**

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Exploration Neural Networks

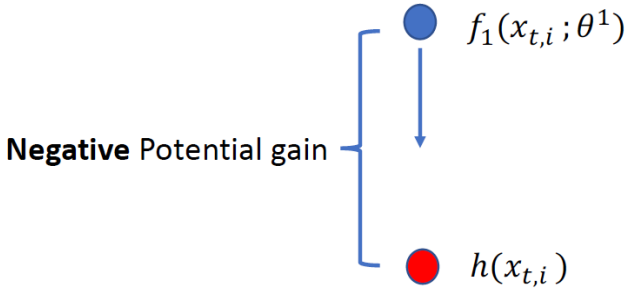
- **Exploration neural network f_2** (novel component) to estimate potential gain:
  - Potential gain has good property: **indicating exploration direction.**

$$h(x_{t,i}) - f_1(x_{t,i}; \theta^1) > 0$$



Case 1: **Upward Exploration**

$$h(x_{t,i}) - f_1(x_{t,i}; \theta^1) < 0$$



Case 2: **Downward Exploration**

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Exploration Neural Networks

- **Exploration neural network f_2**(novel component) to estimate potential gain:
  - Label of f_2: $r_{t,i} - f_1(x_{t,i};)$

$$f_2(x_{t,i};\theta^2) = \boldsymbol{W}_L\sigma(\boldsymbol{W}_{L-1}\sigma(\dots\sigma(\boldsymbol{W}_1\cdot)))$$

  - What is input of f_2?

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Exploration Neural Networks in Bandits

- **Exploration neural network f_2** (novel component) to estimate potential gain:
  - Input of f_2: **Gradient of f_1 with respect to θ^1**:

$$\nabla_{\theta^1} f(x_{t,i}; \theta^1)$$

- **Rational**:
  - Incorporate both **feature of input** and **discriminative information** of f_1.
  - Based on [2,3], f_1 has the following confidence bound:

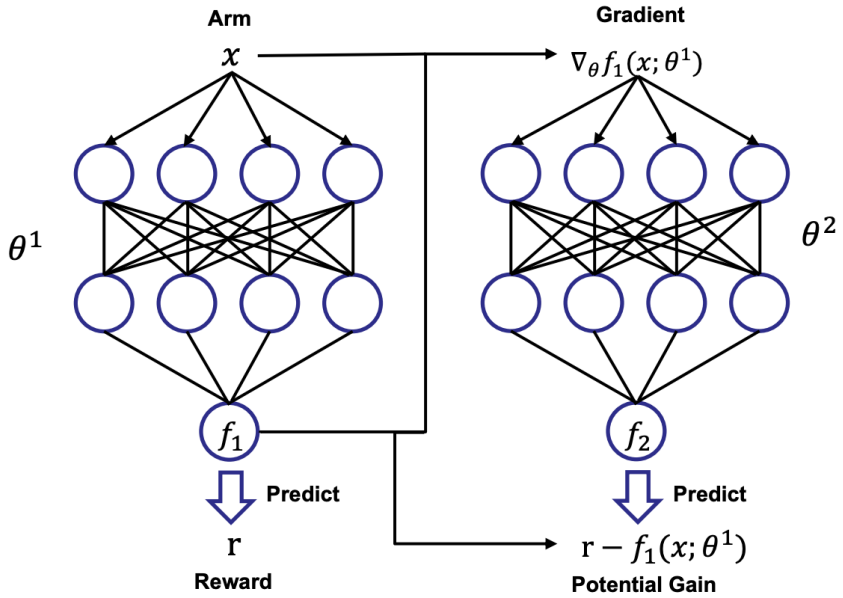
$$|h(\mathbf{x}_{t,i}) - f_1(\mathbf{x}_{t,i}; \boldsymbol{\theta}^1_{t-1})| \leq \Psi(\nabla_{\boldsymbol{\theta}^1_{t-1}} f_1(\mathbf{x}_{t,i}; \boldsymbol{\theta}^1_{t-1})),$$

    Here, instead of choosing a fixed form Ψ, we use f_2 to learn it.

  - In this way, θ^2 is trained on $\{\nabla_{\theta^1} f(x_{t,i}; \theta^1_{\tau-1})\}_{\tau=1}^{t}$ to store historical information.

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

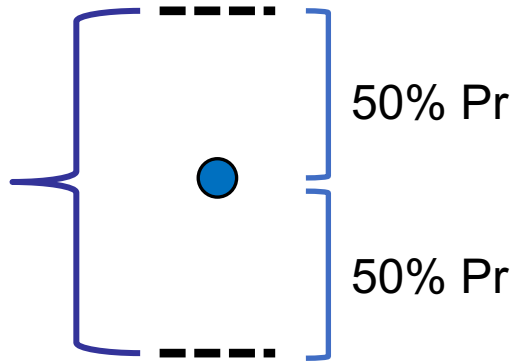
EE-Net: Overview



Exploitation Network f_1 Exploration Network f_2

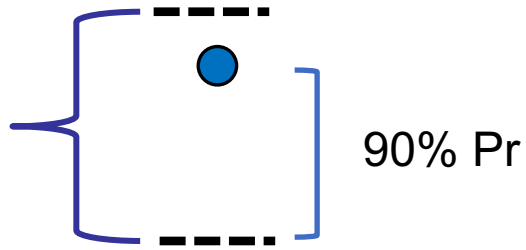
Methods	"Upward" Exploration	"Downward" Exploration
ϵ-Greedy	×	×
NeuralUCB	√	×
NeuralTS	Randomly	Randomly
EE-Net	√	√

Statistical
Confidence Interval



Symmetric and Fixed

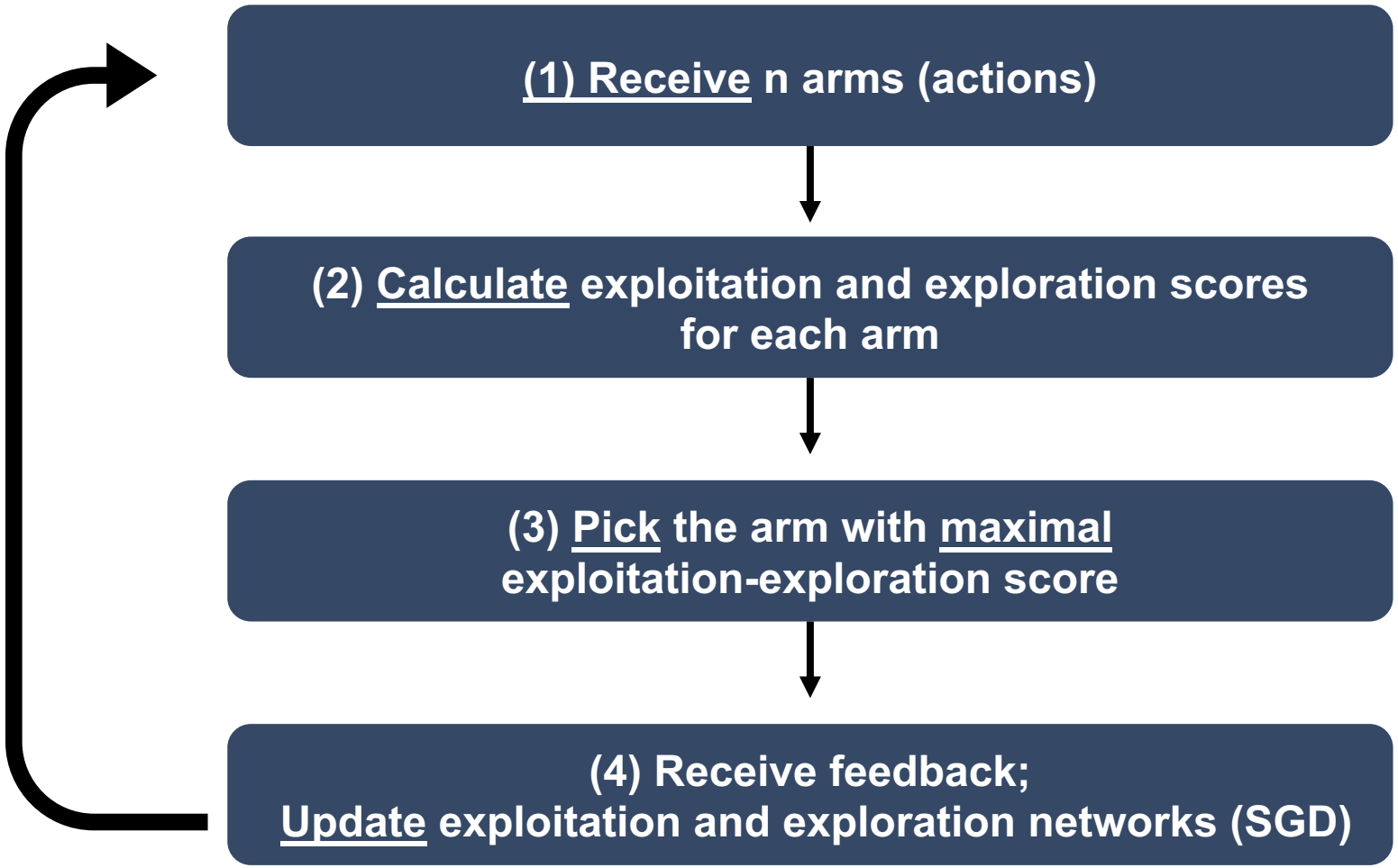
Confidence Interval
learned by neural network
(**Our approach**)



Asymmetric and Adaptive

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Workflow

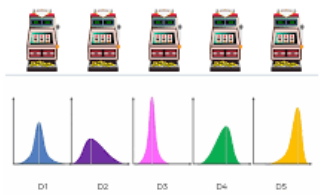


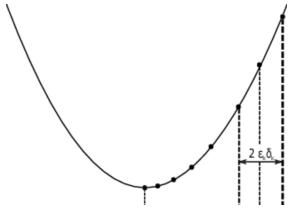
- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

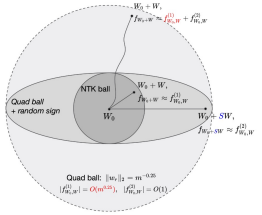
EE-Net: Theoretical Analysis

➢ Proof Workflow of NeuralUCB [1] and NeuralTS [2]:

Gradient Descent in Bandits ⟷ **Gradient Descent on Ridge Regression** ⟷ **NTK Regression** ⟷ **Expected Reward**

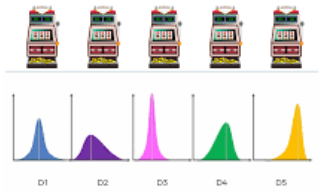


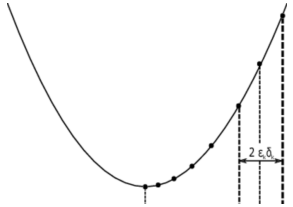




➢ Proof Workflow of EE-Net [3,4]:

Gradient Descent in Bandits ⟷ **Online Gradient Descent** ⟷ **Expected Reward**





[1] Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML, 2020.

[2] Zhang, Weitong, et al. "Neural thompson sampling." ICLR 2021.

[3] Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.

[4] Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Theoretical Analysis

Assumption 1: *For any* $t \in [T], i \in [n], \|\mathbf{x}_{t,i}\|_2 = 1$, *and* $r_{t,i} \in [0,1]$.

- Assumption 1 is standard and mild in analysis of over-parameterized neural networks.
- **No assumption** on distribution of arm contexts.
- Then, we have the following **average error bound for exploration network** f_2:

Lemma1. *For any* $\delta \in (0,1), R > 0$, *suppose* m *satisfies the conditions in Theorem 6. In round* $t \in [T]$, *let*

$$\widehat{i} = \arg\max_{i \in [k]} \left(f_1(\mathbf{x}_{t,\widehat{i}}; \boldsymbol{\theta}_{t-1}^1)/\sqrt{m} + f_2(\phi(\mathbf{x}_{t,\widehat{i}}); \boldsymbol{\theta}_{t-1}^2)/\sqrt{m} \right).$$

Then, with probability at least $1-\delta$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{r_{t,\widehat{i}}} \left[\min\left\{ \left| f_2(\phi(\mathbf{x}_{t,\widehat{i}}); \boldsymbol{\theta}_{t-1}^2)/\sqrt{m} - (r_{t,\widehat{i}} - f_1(\mathbf{x}_{t,\widehat{i}}; \boldsymbol{\theta}_{t-1}^1)/\sqrt{m}) \right|, 1 \right\} \right] \tag{5.3}$$

$$\leq \underbrace{\sqrt{\frac{\Psi(\boldsymbol{\theta}_0^2, R)}{T}}}_{(1)} + \underbrace{\mathcal{O}\left(\frac{3LR}{\sqrt{2T}}\right)}_{(2)} + \underbrace{\sqrt{\frac{2\log(\mathcal{O}(1)/\delta)}{T}}}_{(3)}.$$

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Theoretical Analysis

Lemma 1. *For any $\delta \in (0,1), R > 0$, suppose m satisfies the conditions in Theorem 6. In round $t \in [T]$, let*

$$\widehat{i} = \arg\max_{i \in [k]} \left(f_1(\mathbf{x}_{t,\widehat{i}}; \boldsymbol{\theta}_{t-1}^1)/\sqrt{m} + f_2(\phi(\mathbf{x}_{t,\widehat{i}}); \boldsymbol{\theta}_{t-1}^2)/\sqrt{m} \right).$$

Then, with probability at least $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{r_{t,\widehat{i}}} \left[\min\left\{ \left| f_2(\phi(\mathbf{x}_{t,\widehat{i}}); \boldsymbol{\theta}_{t-1}^2)/\sqrt{m} - (r_{t,\widehat{i}} - f_1(\mathbf{x}_{t,\widehat{i}}; \boldsymbol{\theta}_{t-1}^1)/\sqrt{m}) \right|, 1 \right\} \right] \leq \underbrace{\sqrt{\frac{\Psi(\boldsymbol{\theta}_0^2, R)}{T}}}_{(1)} + \underbrace{\mathcal{O}\left(\frac{3LR}{\sqrt{2T}}\right)}_{(2)} + \underbrace{\sqrt{\frac{2\log(\mathcal{O}(1)/\delta)}{T}}}_{(3)}. \tag{5.3}$$

- **(1)** Complexity term Ψ: **Infimum of regression error** caused by function class $B(\theta^2, R)$:

$$\mathcal{B}(\boldsymbol{\theta}_0^2, R) = \{\widetilde{\boldsymbol{\theta}}^2 \in \mathbb{R}^p : \|\widetilde{\boldsymbol{\theta}}^2 - \boldsymbol{\theta}_0^2\|_2 \leq \mathcal{O}(\frac{R}{\sqrt{m}})\}. \qquad \Psi(\boldsymbol{\theta}_0^2, R) = \inf_{\widetilde{\boldsymbol{\theta}}^2 \in \mathcal{B}(\boldsymbol{\theta}_0^2, R)} \sum_{t=1}^{T} (f^2(\mathbf{x}_{t,\widehat{i}}; \widetilde{\boldsymbol{\theta}}^2) - r_{t,\widehat{i}}^2)^2$$

- **(2) Price** of picking function class $B(\theta^2, R)$ controlled by radius R.
- **(3) Confidence bound** for predictions of f_2.

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

EE-Net: Regret Upper Bound

- Then, we have following regret upper bound $\tilde{O}(\sqrt{T})$ for EE-Net:

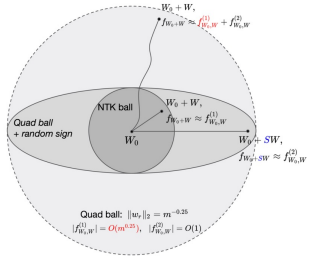
Theorem 1. *Let f_1, f_2 follow the setting of f (Eq. (5.1)) with the same width m and depth L. Suppose $m \geq \Omega\left(poly(T, L, R, \log(1/\delta))\right)$, $\eta_1 = \eta_2 = \frac{\sqrt{\nu}R}{m\sqrt{T}}$ and $\Psi(\boldsymbol{\theta}_0^2, R) \& \Psi^*(\boldsymbol{\theta}_0^2, R) \leq \Psi$. Then, for any $\delta \in (0,1), R > 0$, with probability at least $1-\delta$ over the initialization, there exists a constant ν, such that the pseudo regret of Algorithm 1 in T rounds satisfies*

$$\mathbf{R}_T \leq \sqrt{T} \cdot \mathcal{O}\left(RL + \sqrt{\Psi} + 2\sqrt{2\log(\mathcal{O}(1)/\delta)}\right) + \mathcal{O}(1) \tag{5.2}$$

- Compared to existing works NeuralUCB [3] and NeuralTS [4]:

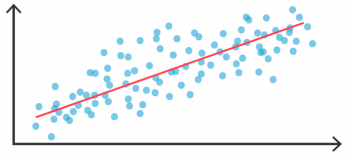
$$\mathbf{R}_T \leq \mathcal{O}\left(\sqrt{\tilde{d}T\log T + S^2}\right) \cdot \mathcal{O}\left(\sqrt{\tilde{d}\log T}\right), \quad \text{and} \quad \tilde{d} = \frac{\log\det(\mathbf{I} + \mathbf{H}/\lambda)}{\log(1 + Tn/\lambda)}$$

NTK

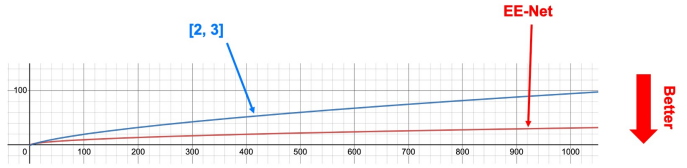


vs

Regression

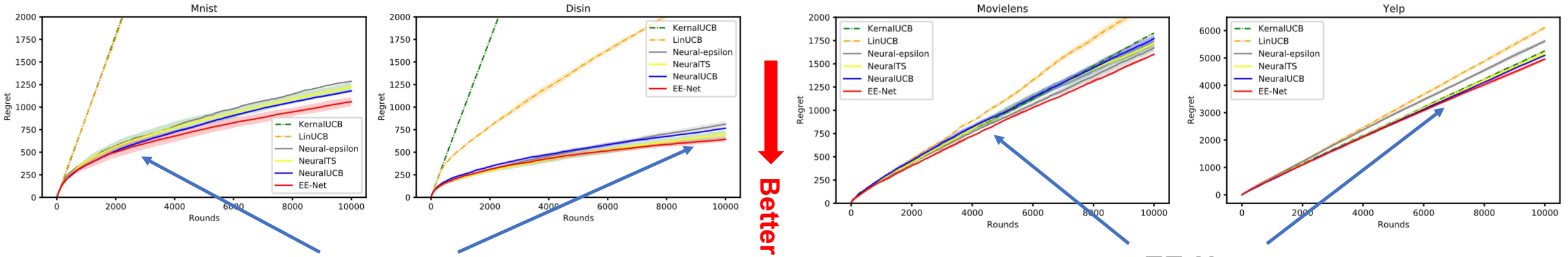


1) **[Better Interpretability]:** Have the similar complexity term but Ψ **easier to interpret**.
2) **[Contexts]:** Allow arm contexts to be **repeatedly observed**.
3) **[Tighter Bound]:** EE-Net **improves by a multiplicative factor** $\log T$.



[1] Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
[2] Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.
[3] Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML, 2020.
[4] Zhang, Weitong, et al. "Neural thompson sampling." ICLR 2021.

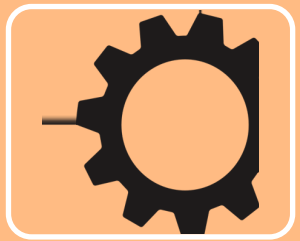
EE-Net: Empirical Experiments



- Setup:
  - Classification and recommendation dataset.
  - 5 state-of-the-art baselines including ϵ-greedy, UCB, TS exploration strategy.
  - All methods have the same exploitation network f_1.
- **EE-Net** achieves **substantial improvements,** because **all improvements purely come from exploration!**

- Ban, Yikun, et al. "EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits." ICLR 2022.
- Ban, Yikun, et al. "Neural Exploitation and Exploration of Contextual Bandits." JMLR 2024.

Neural Contextual Bandits: Roadmap



Fundamental Exploration

- Upper Confidence Bound
- Thompson Sampling
- Exploration Network

Efficient Exploration

- Neural Linear UCB
- Neural Network with Perturbed Reward
- Inverse Weight Gap Strategy

Neural Linear UCB

- In each round, a user is serving

for $t = 1, \ldots, T$ **do**
  receive feature vectors $\{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,K}\}$
  choose arm $a_t = \text{argmax}_{k \in [K]} \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}(\mathbf{x}_{t,k}; \mathbf{w}_{t-1}) + \alpha_t \|\boldsymbol{\phi}(\mathbf{x}_{t,k}; \mathbf{w}_{t-1})\|_{\mathbf{A}_{t-1}^{-1}}$, and obtain reward $\widehat{r}_t$
  update $\mathbf{A}_t$ and $\mathbf{b}_t$ as follows:
  $\mathbf{A}_t = \mathbf{A}_{t-1} + \boldsymbol{\phi}(\mathbf{x}_{t,a_t}; \mathbf{w}_{t-1}) \boldsymbol{\phi}(\mathbf{x}_{t,a_t}; \mathbf{w}_{t-1})^\top$, $\mathbf{b}_t = \mathbf{b}_{t-1} + \widehat{r}_t \boldsymbol{\phi}(\mathbf{x}_{t,a_t}; \mathbf{w}_{t-1})$,
  update $\boldsymbol{\theta}_t = \mathbf{A}_t^{-1} \mathbf{b}_t$
  if $\text{mod}(t, H) = 0$ **then**
    $\mathbf{w}_t \leftarrow$ output of Algorithm 2
    $q = q + 1$
  else
    $\mathbf{w}_t = \mathbf{w}_{t-1}$
  end if
end for
Output $\mathbf{w}_T$

K arms

Representation by Neural Network

Exploitation

Exploration

Estimated by Ridge Regression

Compared with LinUCB (Li et al. 2010)

$$U_{t,a} = \underbrace{\langle \mathbf{x}_{t,a}, \boldsymbol{\theta}_{t-1} \rangle}_{\text{mean}} + \gamma_{t-1} \underbrace{\sqrt{\mathbf{x}_{t,a}^\top \mathbf{Z}_{t-1}^{-1} \mathbf{x}_{t,a}}}_{\text{variance}}$$

$$\boldsymbol{\phi}(\mathbf{x}; \mathbf{w}) = \sqrt{m} \sigma(\mathbf{W}_L \sigma(\mathbf{W}_{L-1} \cdots \sigma(\mathbf{W}_1 \mathbf{x}) \cdots)).$$

- Update Neural Network Parameter:

  Loss function:

$$\mathcal{L}_q(\mathbf{w}) = \sum_{i=1}^{qH} \left(\boldsymbol{\theta}_i^\top \phi(\mathbf{x}_{i,a_i}; \mathbf{w}) - \widehat{r}_i \right)^2.$$

  - Gradient Descent.

- Xu, Pan, et al. “Neural contextual bandits with deep representation and shallow exploration.” ICLR 2022.

Neural Bandit With Perturbed Reward

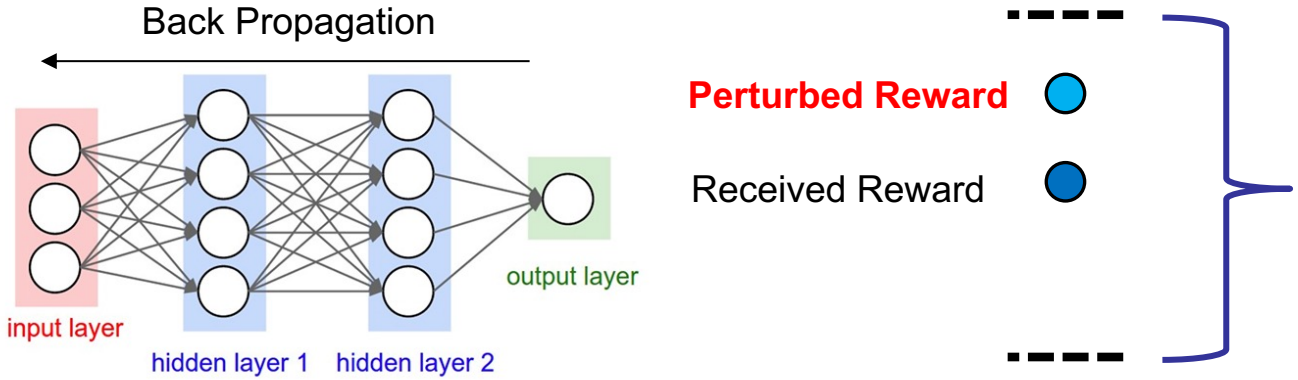
- In each round, a user is serving

for $t = 1, \ldots, T$ **do**
  if $t > K$ **then** Initialization: Pull each arm once
    Pull arm a_t and receive reward r_{t,a_t}, where $a_t = \text{argmax}_{i \in [K]} f(\mathbf{x}_i, \boldsymbol{\theta}_{t-1})$. Selection Criterion
    Generate $\{\gamma_s^t\}_{s \in [t]} \sim \mathcal{N}(0, \nu^2)$. Perturbed Reward
    Set $\boldsymbol{\theta}_t$ by the output of gradient descent for solving Eq (3.2).
  else
    Pull arm a_k.
  end if
end for

Reward Perturbation (Noise)

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{s=1}^{t} \left(f(\mathbf{x}_{a_s}; \boldsymbol{\theta}) - (r_{s,a_s} + \gamma_s^t) \right)^2 / 2 + m\lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 / 2$$

Implicit Exploration:



- Jia, Yiling, Weitong Zhang, Dongruo Zhou, Quanquan Gu, and Hongning Wang. "Learning neural contextual bandits through perturbed rewards." ICLR 2022.

Neural SquareCB: Inverse Gap Strategy

- In each round, a user is serving

Special Case: $y = 1 - r$

Estimated loss by neural networks

for $t = 1, 2, ..., T$ **do** K arms

Receive contexts $\mathbf{x}_{t,1}, ..., \mathbf{x}_{t,K}$, and compute $\hat{y}_{t,a} = \tilde{f}^{(S)}\left(\theta; \mathbf{x}_{t,a}, \boldsymbol{\varepsilon}^{(1:S)}\right), \forall a \in [K]$

Let $b = \arg\min_a \hat{y}_{t,a}$, $p_{t,a} = \frac{1}{K + \gamma(\hat{y}_{t,b} - \hat{y}_{t,a})}$, and $p_{t,b} = 1 - \sum_{a \neq b} p_{t,a}$

Sample arm $a_t \sim p_t$ and observe output y_{t,a_t}

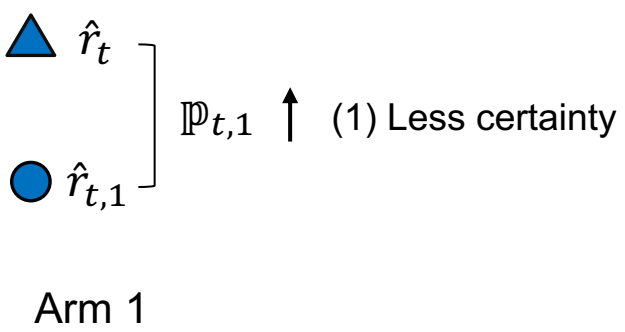
Update $\theta_{t+1} = \prod_{B^{\text{Frob}}_{\rho,\rho_1}(\theta_0)} \left(\theta_t - \eta_t \nabla \mathcal{L}^{(S)}_{\text{Sq}}\left(y_{t,a_t}, \left\{\tilde{f}(\theta; \mathbf{x}_{t,a_t}, \boldsymbol{\varepsilon}_s)\right\}_{s=1}^{S}\right)\right)$.

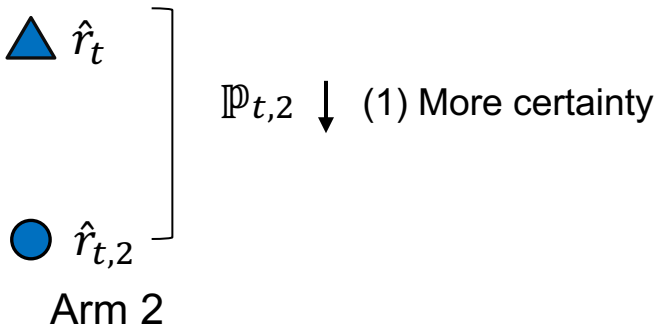
end for

Equal to the **arm with maximal reward**

Inverse Weight Gap to form distribution for Selection

$$\hat{r}_t = \arg\max f(x_{t,i}; \theta_t) \qquad \mathbb{p}_{t,i} \propto \frac{1}{\hat{r}_t - \hat{r}_{t,i}}$$ Selection Probability





• Deb, Rohan, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. "Contextual bandits with online neural regression." ICLR 2024.

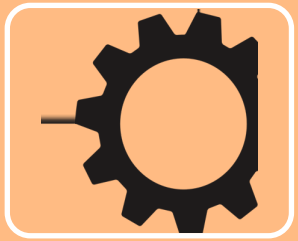
Neural SquareCB: Inverse Gap Strategy

Algorithm	Regret	Remarks
Neural UCB [Zhou et al., 2020]	$\tilde{\mathcal{O}}(\tilde{d}\sqrt{T})$	Bound depends on $\tilde{d}$ and could be $\Omega(T)$ in worst case.
Neural TS Zhang et al. [2021]	$\tilde{\mathcal{O}}(\tilde{d}\sqrt{T})$	Bound depends on $\tilde{d}$ and could be $\Omega(T)$ in worst case.
EE-Net [Ban et al., 2022b]	$\tilde{\mathcal{O}}(\sqrt{T})$	Assumes that the contexts at every round are drawn i.i.d and needs to store all the previous networks.
NeuSquareCB (This work)	$\tilde{\mathcal{O}}(\sqrt{KT})$	No dependence on $\tilde{d}$ and holds even when the contexts are chosen adversarially.
NeuFastCB (This work)	$\tilde{\mathcal{O}}(\sqrt{L^*K} + K)$	No dependence on $\tilde{d}$ and holds even when the contexts are chosen adversarially. Further, this is the first data-dependent regret bound for neural bandits.

- Remove dependence of effective dimension.
- Minimize dependence on Neural Tangent Kernel.

Deb, Rohan, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. "Contextual bandits with online neural regression." ICLR 2024.

Takeaways



Fundamental Exploration

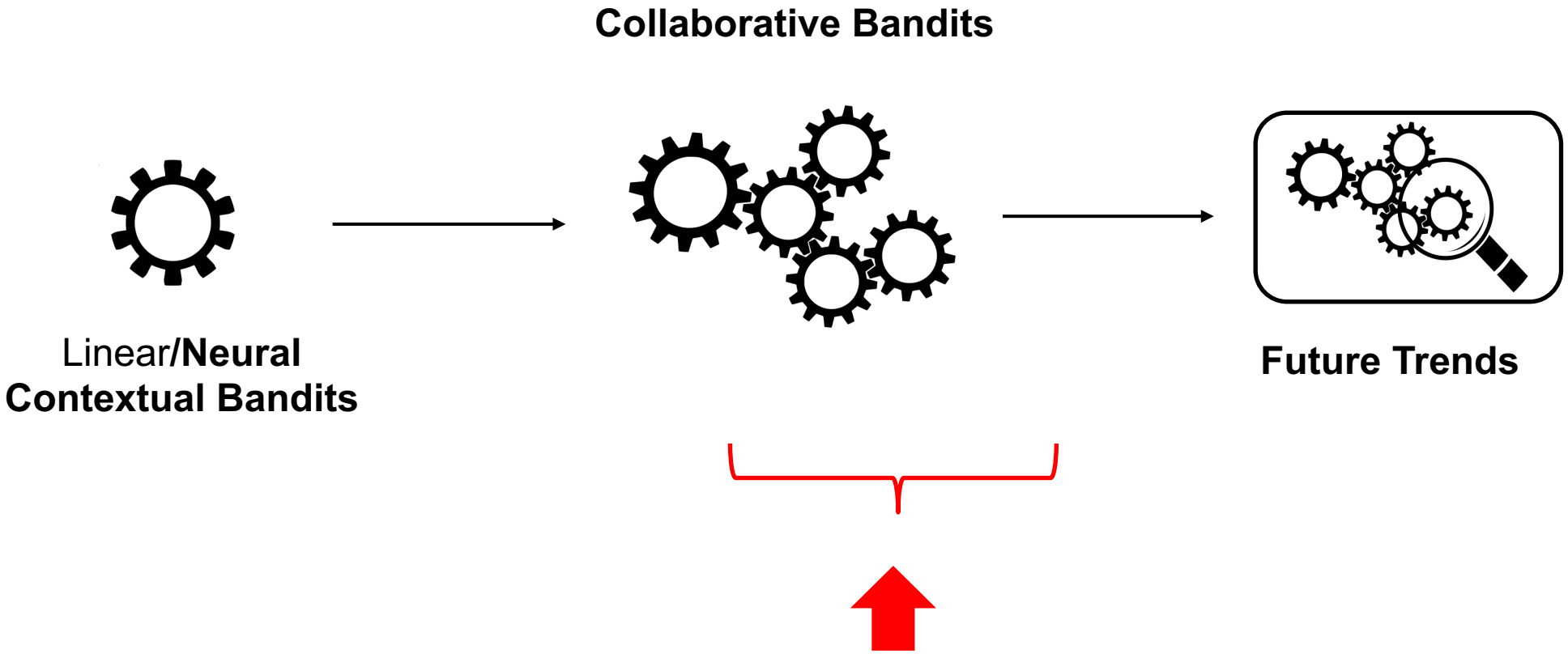
- Neural UCB [1] -- An Extension of LinUCB to NTK Space
- Neural TS [2] -- An Extension of LinTS to NTK Space
- EE-Net [3] -- Another Neural Network for Exploration

Efficient Exploration

- Neural Linear UCB [4] -- LinUCB with Neural Representation
- Neural Network with Perturbed Reward [5] -- Implicit Exploration by Perturbing Rewards
- Neural Square CB[6] -- Exploration using Inverse Weight Gap Strategy

[1] Zhou, Dongruo, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration." ICML 2020.
[2] Zhang, Weitong, Dongruo Zhou, Lihong Li, and Quanquan Gu. "Neural thompson sampling." ICLR 2021.
[3] Ban, Yikun, Yuchen Yan, Arindam Banerjee, and Jingrui He. "Ee-net: Exploitation-exploration neural networks in contextual bandits." ICLR 2022.
[4] Xu, Pan, et al. “Neural contextual bandits with deep representation and shallow exploration.” ICLR 2022.
[5] Jia, Yiling, Weitong Zhang, Dongruo Zhou, Quanquan Gu, and Hongning Wang. "Learning neural contextual bandits through perturbed rewards." ICLR 2022.
[6] Deb, Rohan, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. "Contextual bandits with online neural regression." ICLR 2024.

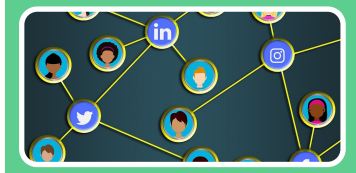
Roadmap


Collaborative Bandits
Linear/Neural
Contextual Bandits
Future Trends

Collaborative Bandits

Introduction
- Background & Motivations
- Challenges

Online Clustering of Bandits
- Clustering of Linear Bandits
- Clustering of Neural Bandits



Graph Bandit Learning with Collaboration
- User side: Graph Neural Bandits
- Arm side: Neural Bandit with Arm Group Graph
- Other Scenarios: Bandit Learning with Graph Feedback & Online Graph Classification with Neural Bandit

Bandits for Combo Recommendation
- Multi-facet Contextual Bandits

Collaborative Contextual Bandits: Background & Motivation

❑ Conventional approaches, e.g., **collaborative and content-based filtering:**

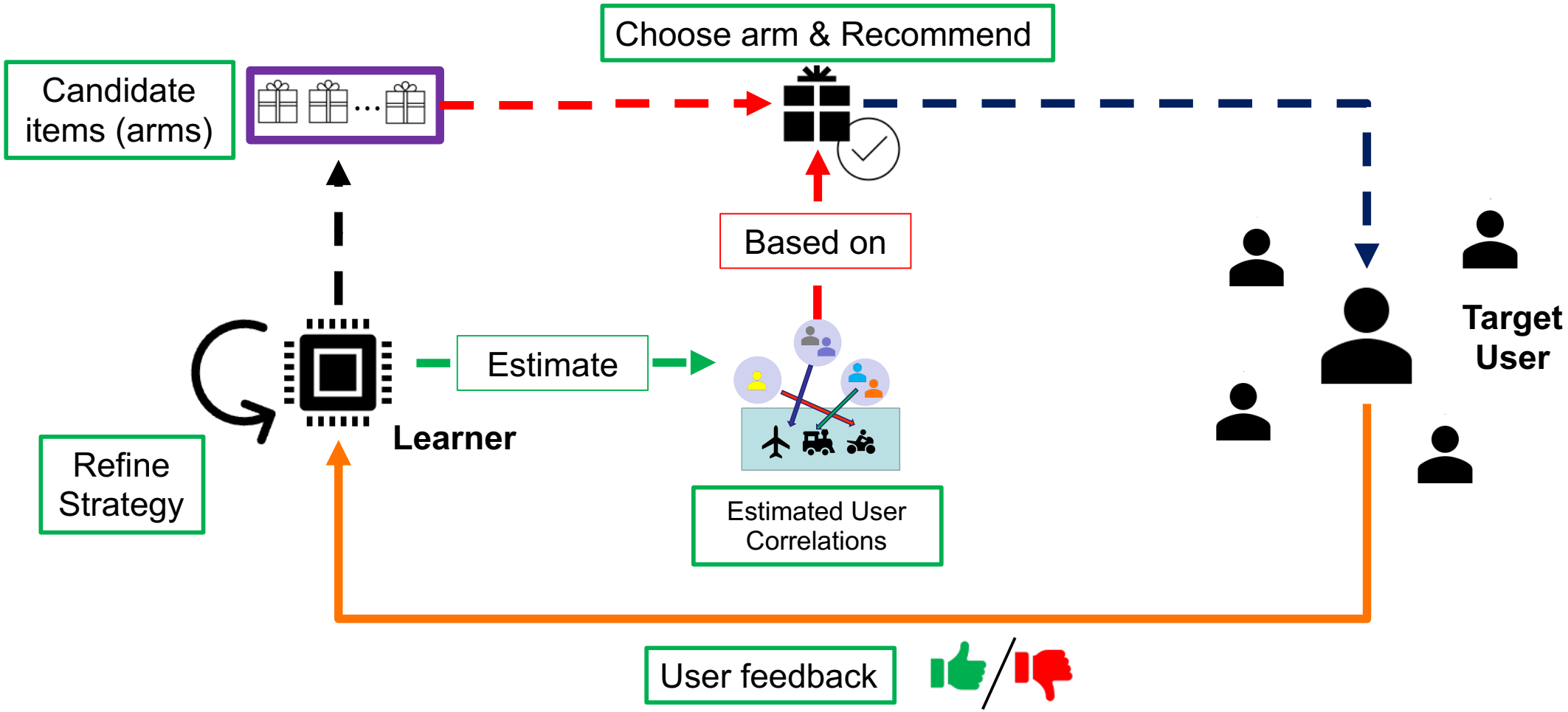
(InCube Group)

Challenges:

- ❑ **Cold-start** problem (Lack of history data);
- ❑ **Rapid change** of recommendation content and user interests.
- ❑ Dilemma of **Exploitation** and **Exploration.**

Collaborative Contextual Bandits: Background & Motivation

❑ **Online recommendation scenario** (in each round):



[1] Lihong Li, et al. 2010. A contextual-bandit approach to personalized news article recommendation. In WWW. 661–670.
[2] Claudio Gentile, et al. 2014. Online clustering of bandits. In ICML. 757–765.

Collaborative Contextual Bandits: Background & Motivation

- **The dilemma of exploitation and exploration:**

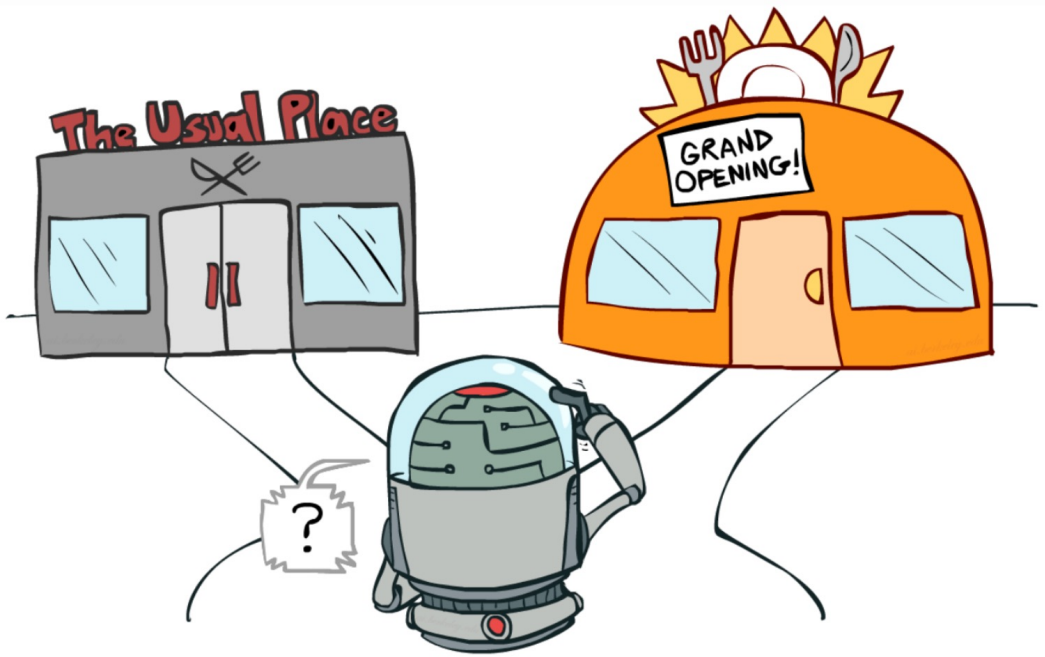
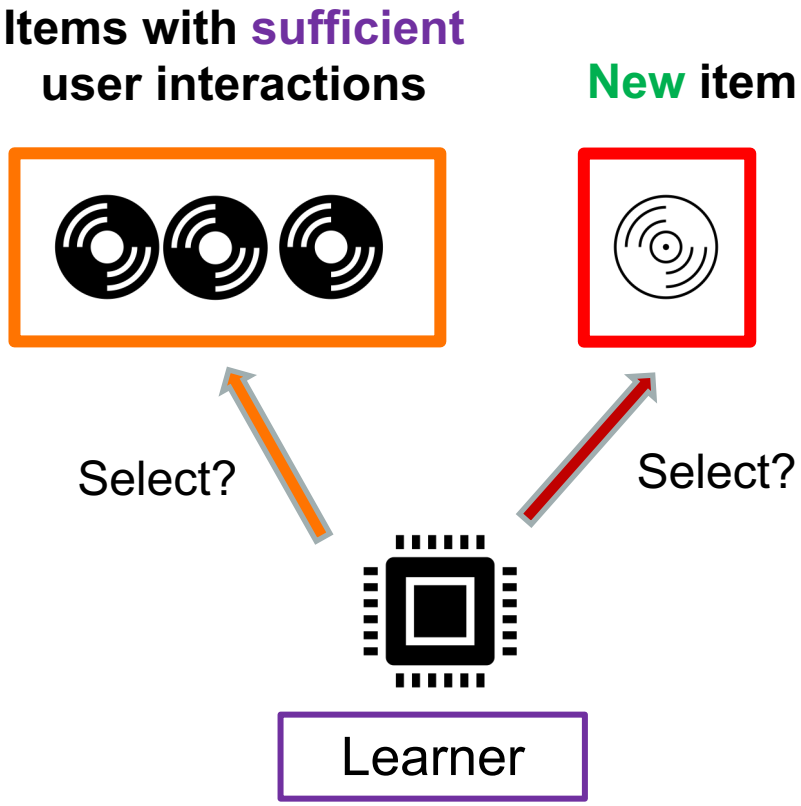
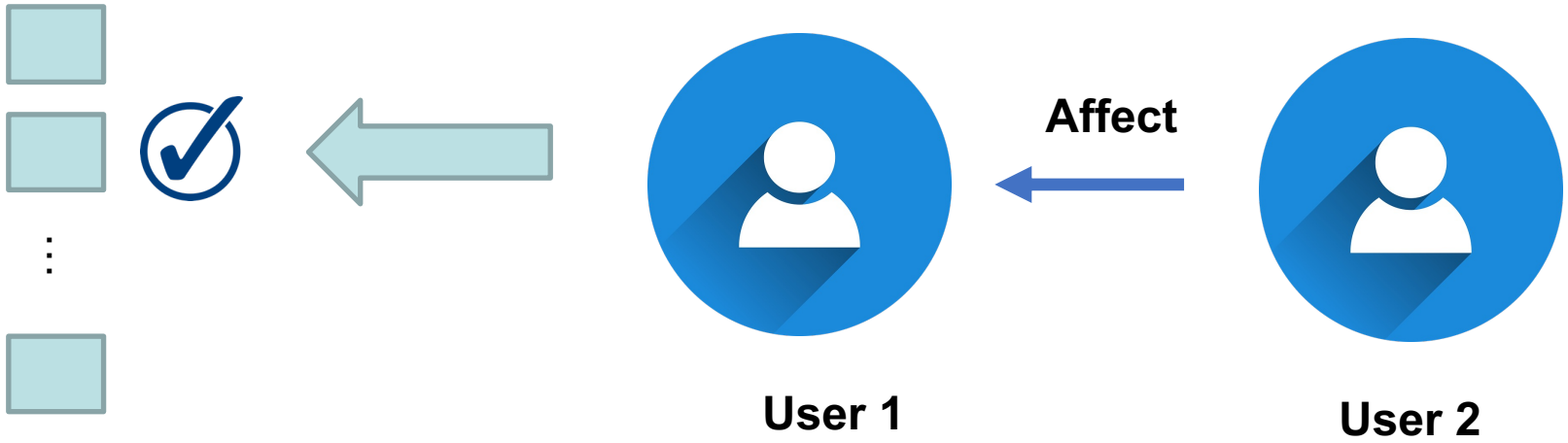


Figure: UC Berkeley CS 188, Introduction to Artificial Intelligence



Collaborative Contextual Bandits: Background & Motivation

❑ **One user's decision is affected by other users.**



❑ **Motivations**: Utilizing the **mutual influence / user collaborative effects** can

- Improve **recommendation quality**.
- Alleviate the **interaction scarcity** issue in terms of individual users.
- Rapidly adapt to **new users / items** based on interactions with other users.

Collaborative Contextual Bandits: Challenges

- ***Challenge* #1**: How to **formally model user collaborations?**
  - User clusters [1, 2, 3, 4, 5, 6, 7], graphs with user nodes [10], etc.
- ***Challenge* #2**: How to **discover user correlations?**
  - Leveraging the **known** user correlation information from the environment [8, 9];
  - User clustering based on their past interactions [2,3,4,5,7], exploitation-exploration graph construction [10].
- ***Challenge* #3**: How to **utilize user correlation** to **improve recommendation quality?**
  - Combination of linear estimations [1, 2, 3, 4, 5, 6], gradient-based meta-learning [7], graph neural networks [10], etc.

[1] Gentile et. al., Online clustering of bandits. ICML 2014.
[2] Li et. al., Improved algorithm on online clustering of bandits. IJCAI 2019.
[3] Nguyen et. al., Dynamic clustering of contextual multi-armed bandits. CIKM 2014.
[4] Gentile et. al., On context-dependent clustering of bandits. ICML 2017.
[5] Ban et. al., Local clustering in contextual multi-armed bandits. WWW 2021.
[6] Li et. al., Collaborative filtering bandits. SIGIR 2016.
[7] Ban et. al., Meta clustering of neural bandits. In submission.
[8] Nicolo Cesa-Bianch et. al., A gang of bandits. NIPS 2013.
[9] Wu et. al., Contextual bandits in a collaborative environment. SIGIR 2016.
[10] Qi et. al., Graph neural bandits. KDD 2023.

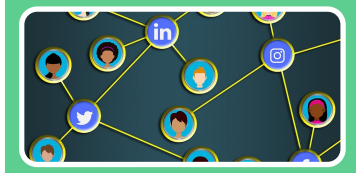
Roadmap

Introduction

- Background & Motivations
- Challenges

Online Clustering of Bandits

- Clustering of Linear Bandits
- Clustering of Neural Bandits



Graph Bandit Learning with Collaboration

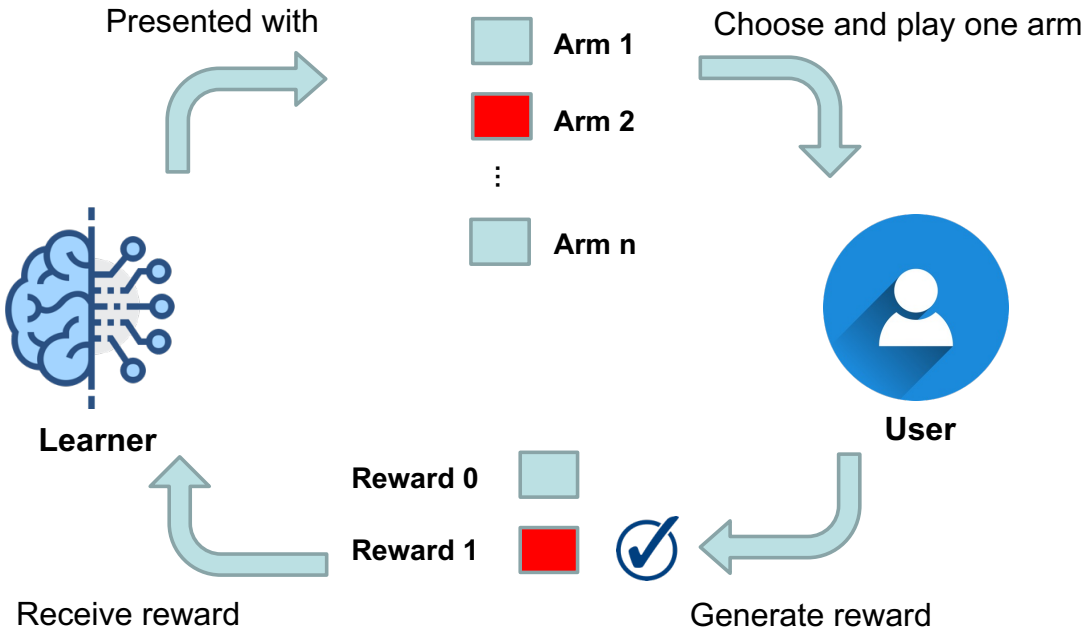
- User side: Graph Neural Bandits
- Arm side: Neural Bandit with Arm Group Graph
- Other Scenarios: Bandit Learning with Graph Feedback & Online Graph Classification with Neural Bandit

Bandits for Combo Recommendation

- Multi-facet Contextual Bandits

Online Clustering of Bandits

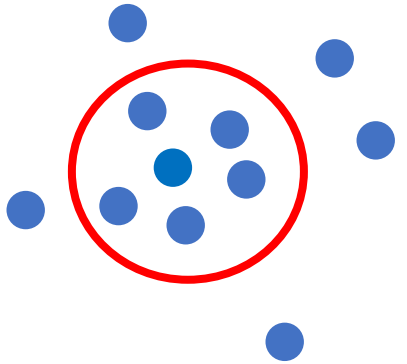
❑ Standard MAB algorithms view each user as an individual, without **user dependency**.



❑ For **refined** recommendation strategies when user correlations are unknown:

- Objective #1: **Identify user clusters** in MAB;
- Objective #2: **Exploit the user clusters** to improve the recommendation.

Online Clustering of Linear Bandits

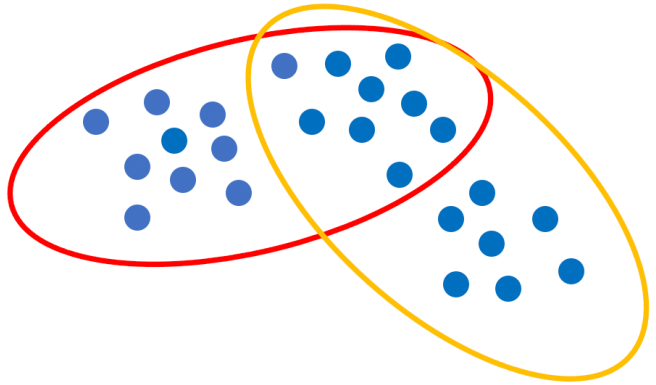


- **Clustering of <u>Linear</u> Bandits:**
  - Under **linear** stochastic contextual bandit settings: $r = \langle \theta_u, x \rangle + \eta$.
  - User **correlation intensity** between u, u' is measured by $\|\theta_u - \theta_{u'}\|_2$.
    1. User clusters with identical preferences [1, 2, 3, 4, 6] ($\forall u, u' \in \mathcal{N}$: $\theta_u = \theta_{u'}$).
    2. **A generalized formulation:** γ-cluster of users [5] ($\forall u, u' \in \mathcal{N}$: $\| \theta_u - \theta_{u'} \|_2 \leq \gamma$).

[1] Gentile et. al., Online clustering of bandits. ICML 2014.
[2] Li et. al., Improved algorithm on online clustering of bandits. IJCAI 2019.
[3] Nguyen et. al., Dynamic clustering of contextual multi-armed bandits. CIKM 2014.
[4] Gentile et. al., On context-dependent clustering of bandits. ICML 2017.
[5] Ban et. al., Local clustering in contextual multi-armed bandits. WWW 2021.
[6] Li et. al., Collaborative filtering bandits. SIGIR 2016.

LOCB[1]: Motivation and Challenges

- **Challenge 1: When to ensure a set of identified users is a true cluster?**
  - Cluster: A set of users with similar expected rewards.
  - Expected rewards of users are unknown.



- **Challenge 2: Can we further reduce the clustering complexity?**
  - Previous works have clustering complexity $O(n)$.
  - n is the number of users.

- **Challenge 3: Can we consider and address soft clustering?**
  - Consider overlapping clusters.
  - A user is allowed to belong to multiple clusters.

• Yikun Ban and Jingrui He. Local Clustering in Contextual Multi-Armed Bandits. WWW 2021.

LOCB: Local Clustering of Linear bandits

- **Characterizing similar users' behaviors:**
  - **Definition (γ-Cluster)**: Given a subset of users $\mathcal{N} \subseteq N$ and a threshold $\gamma > 0$, $\mathcal{N}$ is considered a γ-Cluster if it satisfies: $\forall i, j \in \mathcal{N}, \left\|\theta^i - \theta^j\right\| < \gamma$.

- **Objectives:**
  - **Objective #1**: **Identify clusters** among users, such that the clusters returned by the proposed algorithm are true γ–Clusters with probability at least $1-\delta$.
  - **Objective #2**: Leverage user clusters to improve the quality of recommendation, evaluated by **Regret**.

$$\mathbf{R}_T = \mathbb{E}[\sum_{t=1}^{T} R_t] = \sum_{t=1}^{T} (\boldsymbol{\theta}_{i_t}^{\mathsf{T}} \mathbf{x}_t^* - \boldsymbol{\theta}_{i_t}^{\mathsf{T}} \mathbf{x}_t)$$

Optimal Reward

Received Reward

- **Clustering Module + Pulling Module**

• Yikun Ban and Jingrui He. Local Clustering in Contextual Multi-Armed Bandits. WWW 2021.

LOCB: Clustering Module

- **Identify k clusters, given k seeds in each round**:
  - **Seed selection**: Randomly choose k users.
  - **Neighbors**: Two users are neighbors if they belong to the same γ-cluster.
  - **Potential neighbors**: User i is considered as the potential neighbor of seed user s, when:

$$\|\hat{\theta}_{i,t} - \hat{\theta}_{s,t}\| \le B_{\theta,i}(m_{i,t}, \delta') + B_{\theta,s}(m_{s,t}, \delta').$$

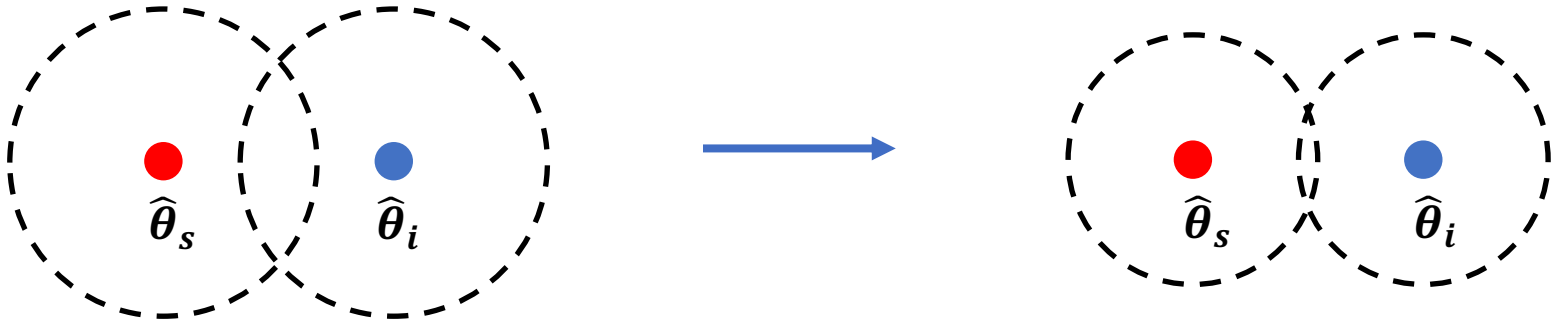
Seed-user parameter User-specific bound Seed-specific bound

$$B_{\theta,i}(m_{i,t}, \delta') = \frac{\sigma\sqrt{2d\log t + 2\log(2/\delta') + 1}}{\sqrt{1 + h(m_{i,t}, H)}},$$

$$h(m_{i,t}, H) = \left(\frac{\lambda m_{i,t}}{4} - 8\log\left(\frac{m_{i,t}+3}{H}\right) - 2\sqrt{m_{i,t}\log\left(\frac{m_{i,t}+3}{H}\right)}\right)$$

  - **Cluster**: Seed user + Its potential neighbors.

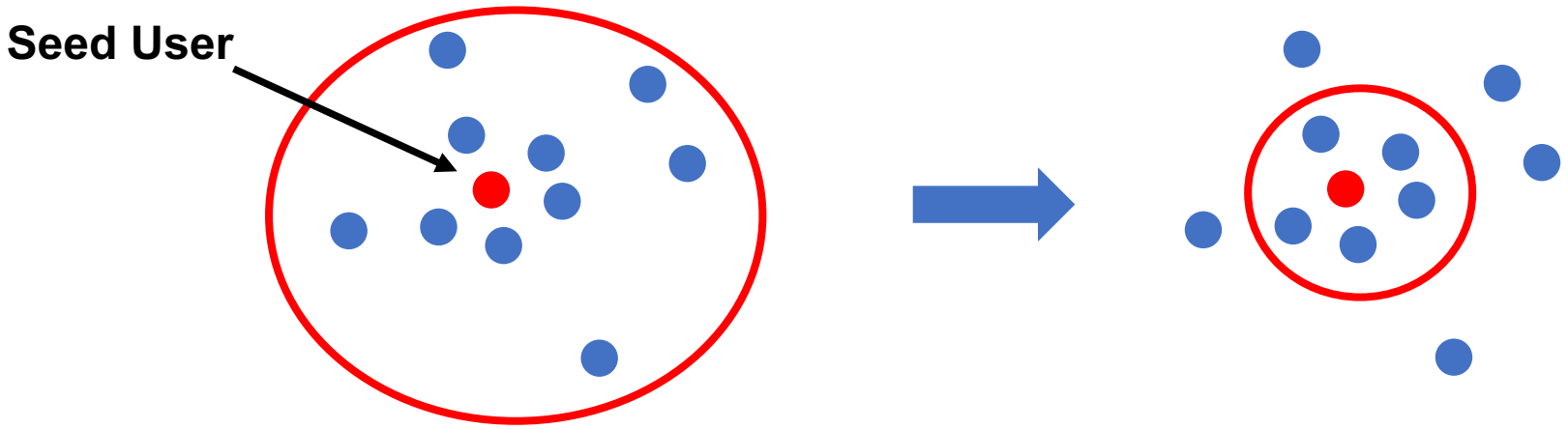
- **User specific bound**: with a high probability, $\left\|\hat{\theta}_{i,t} - \theta_i\right\| \le B_{\theta,i}\left(m_{i,t}, \delta'\right)$



Potential Neighbors

• Yikun Ban and Jingrui He. Local Clustering in Contextual Multi-Armed Bandits. WWW 2021.

LOCB: Evolution of Clusters



Cluster s in Round t Cluster s in Round $t + t'$

- **Evolution of neighbors:** $\|\hat{\theta}_{i,t} - \hat{\theta}_{s,t}\| \leq B_{\theta,i}(m_{i,t}, \delta') + B_{\theta,s}(m_{s,t}, \delta').$

User/seed specific bound is shrinking as more rounds are played for these users.

- **Termination criterion**
  - Given cluster $\mathcal{N}_{s,t}$, Clustering Module outputs this cluster when

$$\sup\{B_{\theta,i}(m_{i,t}, \delta') : i \in \mathcal{N}_{s,t}\} < \frac{\gamma}{8}$$

• Yikun Ban and Jingrui He. Local Clustering in Contextual Multi-Armed Bandits. WWW 2021.

LOCB: Pulling Module

- **Individual CB vs. Cluster CB**
  - Confidence interval for each **cluster**
  - Confidence interval for each **user**

$$\mathbb{P}\left(\forall t \in [T], |\hat{\theta}_{\mathcal{N}_{s,t}}^{\intercal}\mathbf{x}_{a,t} - \theta_{\mathcal{N}_{s,t}}^{\intercal}\mathbf{x}_{a,t}| > CB_{r,\mathcal{N}_{s,t}}\right) < \delta' \qquad \mathbb{P}\left(\forall t \in [T], |\hat{\theta}_{i,t}^{\intercal}\mathbf{x}_{a,t} - \theta_{i}^{\intercal}\mathbf{x}_{a,t}| > CB_{r,i}\right) < \delta'$$

Cluster CB — Individual CB

$$CB_{r,\mathcal{N}_{s,t}} = \frac{1}{|\mathcal{N}_{s,t}|}\sum_{i\in\mathcal{N}_{s,t}} CB_{r,i}$$

- **Pulling Module selects one arm by Cluster UCB**:

$$\mathbf{x}_t = \arg\max_{\mathbf{x}_{a,t}\in\mathbf{X}_t} \hat{\theta}_{\mathcal{N}_{s,t}}^{\intercal}\mathbf{x}_{a,t} + CB_{r,\mathcal{N}_{s,t}}$$

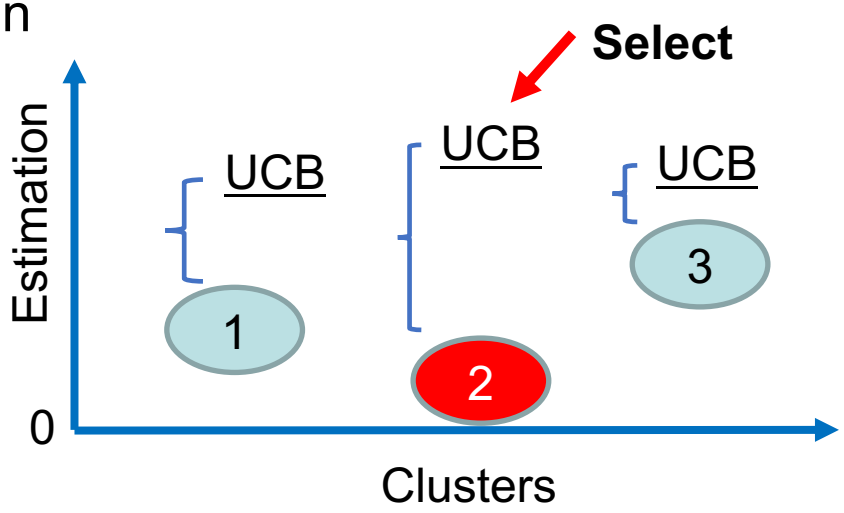
Cluster-level exploration

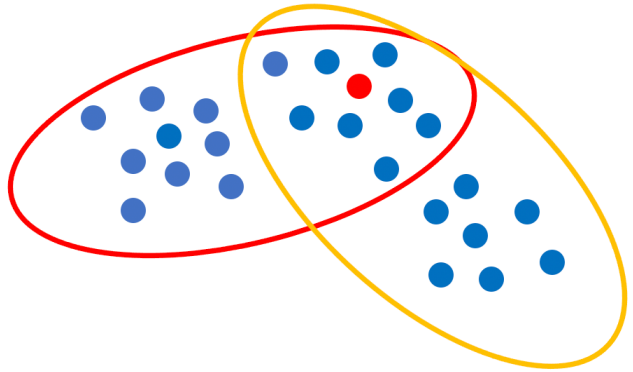
Cluster behavior $\hat{\theta}_{\mathcal{N}_{s,t}} = \frac{1}{|\mathcal{N}_{s,t}|}\sum_{i\in\mathcal{N}_{s,t}} \hat{\theta}_{i,t}.$

- Yikun Ban and Jingrui He. Local Clustering in Contextual Multi-Armed Bandits. WWW 2021.

LOCB: Overlapping Clusters

- **A user may belong to multiple overlapping clusters:**
  - Cluster selection





- Pulling Module selects the **cluster with the maximum potential:**

$$\mathbf{x}_t = \arg \max_{\mathbf{x}_{a,t} \in \mathbf{X}_t} \max_{s \in S_t(i_t)} \left(\hat{\boldsymbol{\theta}}_{\mathcal{N}_{s,t}}^{\intercal} \mathbf{x}_{a,t} + CB_{r,\mathcal{N}_{s,t}} \right)$$

Arm set **Cluster set ($O(k)$)**

• Yikun Ban and Jingrui He. Local Clustering in Contextual Multi-Armed Bandits. WWW 2021.

LOCB: Results

➢ Theoretical analysis:

❑ Correctness ✓

THEOREM 5.1 (CORRECTNESS). *Given a threshold γ and a set of seeds $S \subseteq N$, for each $s \in S$, let $\mathcal{N}_s$ represent the cluster output by LOCB with respect to s. The terminate criterion of Clustering module is defined as:*

$$\sup\{B_{\theta,i}(m_{i,t}, \delta') : i \in \mathcal{N}_{s,t}\} < \frac{\gamma}{8}.$$

Then, with probability at least $1-\delta$, after the Clustering module terminates, for each $s \in S$, it has

$$\forall i, j \in \mathcal{N}_s, \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| < \gamma.$$

❑ Efficiency ✓

THEOREM 5.2. *Suppose each user is evenly served and $m_{i,t} \geq \frac{2\times 32^2}{\lambda^2}\log\left(\frac{2nd}{\delta'}\right)\log\left(\frac{32^2}{\lambda^2}\log\left(\frac{2nd}{\delta'}\right)\right)$ for any $i \in N$. Then, with probability at least $1-\delta$, the number of rounds $\hat{T}$ needed for the Clustering module to terminate is upper bounded by*

$$\hat{T} < \frac{2nd}{C}\log\frac{nd}{C} + \frac{2n}{C}\left(\log(\frac{2^{(d+1)}n}{\delta}) - \frac{\gamma^2 - 256}{512\sigma^2}\right) + n.$$

where $C = \frac{\lambda\gamma^2}{16^3\sigma^2}$.

❑ Effectiveness ✓

THEOREM 5.3. *Suppose that each user is evenly served. Given γ and a set of seeds S, after $T > \hat{T}$ rounds, the accumulated regret of LOCB can be upper bounded as follows:*

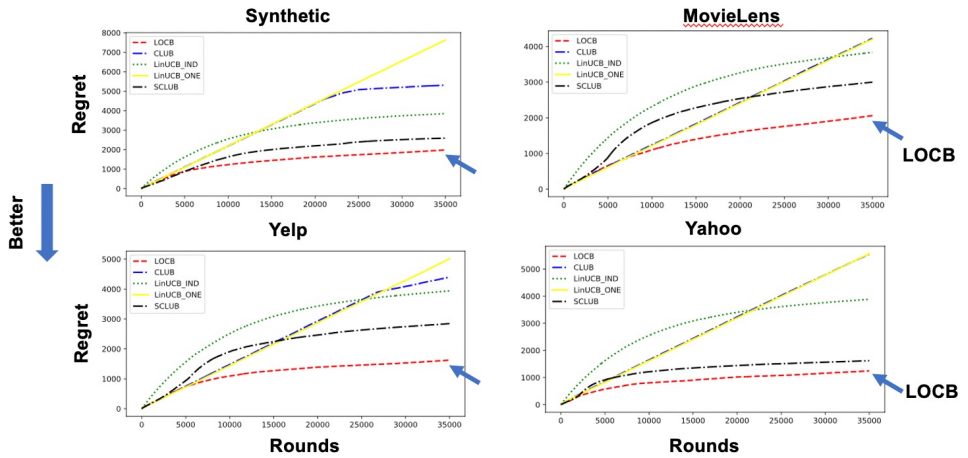
$$\mathbf{R}_T \leq \left[\sqrt{nT}\cdot\sqrt{2d\log(1+T/dn)}\cdot O\left(\sqrt{d\log\left(T/\delta\right)}\right)\right] + \left(T - O\left(nd\log nd\right)\right)\gamma + O\left(nd\log nd\right)\cdot O\left(\sqrt{d\log\left(Tn/\delta\right)}\right).$$

➢ Evaluations:

❑ Improve performance **up to 12.4%.**

	Synthetic			Yelp		
	F1	Pre	Recall	F1	Pre	Recall
N-CLUB	0.390	0.246	0.943	0.484	0.334	0.884
ST-CLUB	0.578	0.549	0.612	0.626	0.593	0.663
ST-SCLUB	0.714	0.745	0.687	0.768	0.863	0.693
N-LOCB	0.662	0.618	0.714	0.675	0.620	0.743
LOCB	**0.880**	**0.913**	**0.856**	**0.879**	**0.908**	**0.853**

	MovieLens			Yahoo		
	F1	Pre	Recall	F1	Pre	Recall
N-CLUB	0.417	0.286	0.773	0.454	0.334	0.709
ST-CLUB	0.520	0.429	0.663	0.528	0.385	0.841
ST-SCLUB	0.538	0.739	0.424	0.632	0.781	0.532
N-LOCB	0.472	0.432	0.524	0.615	0.553	0.692
LOCB	**0.814**	**0.892**	**0.749**	**0.869**	**0.935**	**0.813**



• Yikun Ban and Jingrui He. Local Clustering in Contextual Multi-Armed Bandits. WWW 2021.

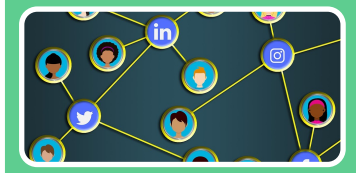
Roadmap

Introduction

- Background & Motivations
- Challenges

Online Clustering of Bandits

- Clustering of Linear Bandits
- Clustering of Neural Bandits



Graph Bandit Learning with Collaboration

- User side: Graph Neural Bandits
- Arm side: Neural Bandit with Arm Group Graph
- Other Scenarios: Bandit Learning with Graph Feedback & Online Graph Classification with Neural Bandit

Bandits for Combo Recommendation

- Multi-facet Contextual Bandits

Online Clustering of Neural Bandits

- **Challenge 1: How to efficiently determining a user's relative group?**
  - User relative group: A set of users with **same expected rewards on a specific item (arm)**.
  - Expected rewards of users are unknown. The mapping function $h(x)$ can be linear or non-linear.

- **Challenge 2: Effective parametric representation of dynamic clusters?**
  - Introducing **meta-learner** capable of representing and swiftly adapting to evolving user clusters.
  - Enabling the rapid acquisition of nonlinear cluster representations.

- **Challenge 3: Balancing exploitation and exploration?**
  - A novel UCB-type exploration strategy.
  - Taking both user-side and meta-side information into account.

• Yikun Ban et. al., Meta Clustering of Neural Bandits. In submission.

M-CNB: Meta Clustering of Neural Bandits

- **Characterizing user clusters without linear assumptions:**

Definition 3.1 (Relative Cluster). In round t, given an arm $\mathbf{x}_{t,i} \in \mathbf{X}_t$, a relative cluster $\mathcal{N}(\mathbf{x}_{t,i}) \subseteq N$ with respect to $\mathbf{x}_{t,i}$ satisfies

(1) $\forall u, u' \in \mathcal{N}(\mathbf{x}_{t,i}), \mathbb{E}[r_{t,i}|u] = \mathbb{E}[r_{t,i}|u']$

(2) $\nexists\, \mathcal{N}' \subseteq N$, s.t. $\mathcal{N}'$ satisfies (1) and $\mathcal{N}(\mathbf{x}_{t,i}) \subset \mathcal{N}'$.

Definition 3.2 (γ-gap). Given two different cluster $\mathcal{N}(\mathbf{x}_{t,i}), \mathcal{N}'(\mathbf{x}_{t,i})$, there exists a constant $\gamma > 0$, such that

$$\forall u \in \mathcal{N}(\mathbf{x}_{t,i}), u' \in \mathcal{N}'(\mathbf{x}_{t,i}), |\mathbb{E}[r_{t,i}|u] - \mathbb{E}[r_{t,i}|u']| \geq \gamma.$$

- **Objectives:**
  - **Objective #1**: **Identify clusters** among users, such that the clusters returned by the proposed algorithm are accurate user clusters.
  - **Objective #2**: Leverage user correlations to improve the quality of recommendation, evaluated by **Pseudo Regret**.

$$\mathbf{R}_T = \sum_{t=1}^{T} \mathbb{E}[r_t^* - r_t \mid u_t, \mathbf{X}_t],$$

Optimal Reward — Received Reward

$$\mathbb{E}[r_t^*|u_t, \mathbf{X}_t] = \max_{\mathbf{x}_{t,i} \in \mathbf{X}_t} h_{u_t}(\mathbf{x}_{t,i})$$

General reward function

• Yikun Ban et. al., Meta Clustering of Neural Bandits. In submission.

M-CNB: Clustering Module

- **Identify relative cluster for target user $u_t \in N$:**
  - **Arm-specific**: Different arms can induce distinct user clusters.
  - **User models**: Each user $u \in N$ is assigned with their own user models $f(\cdot\,;\theta^u)$.
  - **Potential neighbors**: User u is the potential neighbor of target user u_t, when:

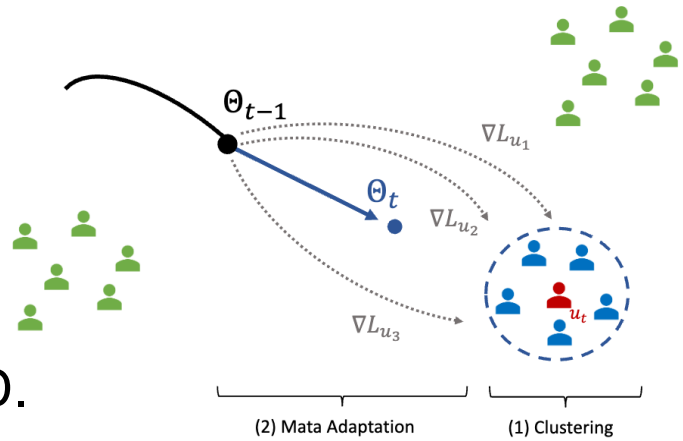
$$\widehat{\mathcal{N}}_{u_t}(\mathbf{x}_{t,i}) = \left\{u \in N \,\middle|\, |f(\mathbf{x}_{t,i};\theta^u_{t-1}) - f(\mathbf{x}_{t,i};\theta^{u_t}_{t-1})| \le \frac{\nu-1}{\nu}\gamma\right\}.$$

Preference est. for other users

Preference est. for target user

Tunable distance threshold

- **Meta-adaptation:** Adapting to estimated user clusters.
  - Randomly draw a few samples from the historical data of detected cluster $\{\mathcal{T}^u_{t-1}\}_{u\in\widehat{\mathcal{N}}_{u_t}(\mathrm{x}_{i,t})}$.
  - The meta-model $f(\cdot\,;\Theta)$ is adapted through a few steps of SGD.



• Yikun Ban et. al., Meta Clustering of Neural Bandits. In submission.

M-CNB: Pulling Module

- **Informative UCB for reward estimation:**

Meta-Model Error

$$\sum_{t=1}^{T} \mathbb{E}_{r_t|\mathbf{x}_t}\left[\left|f(\mathbf{x}_t;\Theta_t) - r_t\right| \,\middle|\, u_t\right]$$

Gradient Discrepancy between User Model and the Meta-Model

$$\leq \sum_{t=1}^{T} \underbrace{\frac{O(\|\nabla_\Theta f(\mathbf{x}_t;\Theta_t) - \nabla_\theta f(\mathbf{x}_t;\theta_0^{u_t})\|_2)}{m^{1/4}}}_{\text{Meta-side info}} + \sum_{u\in N} \mu_T^u \Bigg[\underbrace{O\left(\sqrt{\frac{S+1}{2\mu_T^u}}\right) + \sqrt{\frac{2\log(1/\delta)}{\mu_T^u}}}_{\text{User-side info}} \Bigg],$$

User-side Upper Bound based on Service Frequency

- **Pulling Module selects one arm by Cluster UCB**:

$$\mathbf{x}_t = \arg_{\mathbf{x}_{t,i}\in \mathbf{X}_t} \max \mathbf{U}_{t,i}$$

$$\mathbf{U}_{t,i} = f(\mathbf{x}_{t,i};\Theta_{t,i}) + \frac{\|\nabla_\Theta f(\mathbf{x}_{t,i};\Theta_{t,i}) - \nabla_\theta f(\mathbf{x}_{t,i};\theta_0^{u_t})\|_2}{m^{1/4}} + \sqrt{\frac{S+1}{2\mu_t^u}} + \sqrt{\frac{2\log(1/\delta)}{\mu_t^u}}$$

Meta-Model Reward Estimation

UCB

• Yikun Ban et. al., Meta Clustering of Neural Bandits. In submission.

M-CNB: Theoretical and Empirical Results

➢ Theoretical analysis from two aspects:

❑ Instance-dependent Regret Bound ✓

Theorem 5.1. *Given the number of rounds T and γ, for any $\delta \in (0,1), R > 0$, suppose $m \geq \widetilde{\Omega}(poly(T,L,R) \cdot Kn\log(1/\delta))$, $\eta_1 = \eta_2 = \frac{R^2}{\sqrt{m}}$, and $\mathbb{E}[|\mathcal{N}_{u_t}(\mathbf{x}_t)|] = \frac{n}{q}, t \in [T]$. Then, with probability at least $1-\delta$ over the initialization, Algorithm 1 achieves the following regret upper bound:*

$$\mathbf{R}_T \leq \sqrt{qT \cdot S^*_{TK} + O(1)} + O(\sqrt{2qT\log(O(1)/\delta)}).$$

where $S^*_{TK} = \inf_{\theta \in B(\theta_0,R)} \sum_{t=1}^{TK} \mathcal{L}_t(\theta)$.

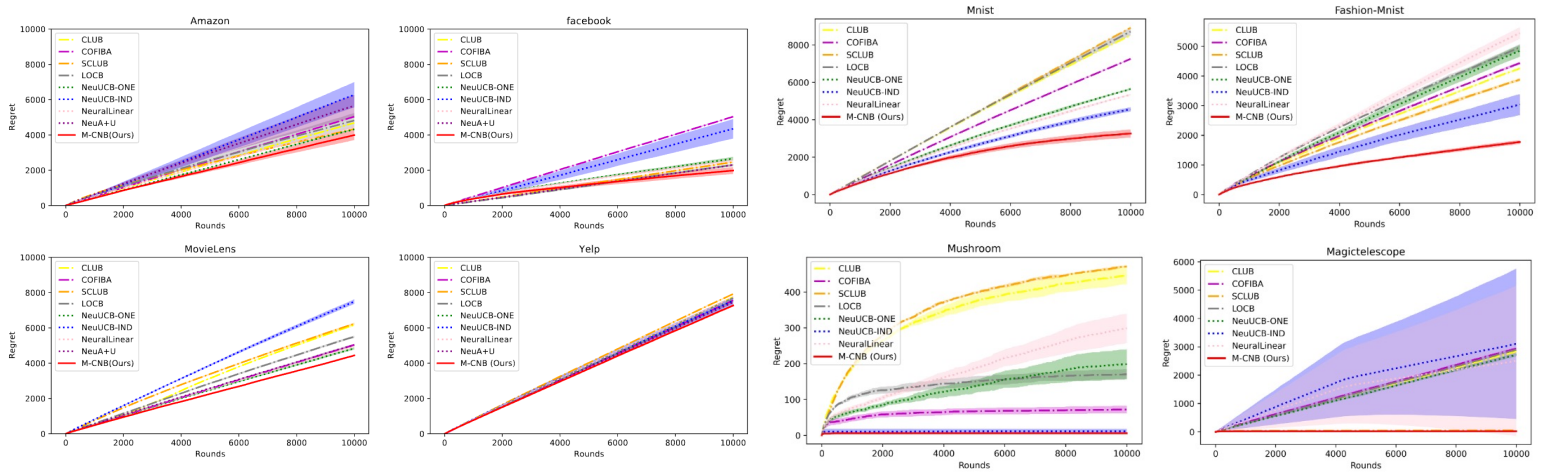
❑ NTK-regression based Regret Bound ✓

Lemma 5.3. *Suppose Assumption 5.1 and conditions in Theorem 5.1 holds where $m \geq \widetilde{\Omega}(poly(T,L) \cdot Kn\lambda_0^{-1}\log(1/\delta))$. With probability at least $1-\delta$ over the initialization, there exists $\theta' \in B(\theta_0, \widetilde{\Omega}(T^{3/2}))$, such that*

$$\mathbb{E}[S^*_{TK}] \leq \mathbb{E}[\sum_{t=1}^{TK} \mathcal{L}_t(\theta')] \leq \widetilde{O}\left(\sqrt{\widetilde{d}} + S\right)^2 \cdot \widetilde{d}.$$

➢ Evaluations:

- ❑ M-CNB (red curve) outperforms baselines, for both recommendation and classification data sets.



• Yikun Ban et. al., Meta Clustering of Neural Bandits. In submission.

Online Clustering of Bandits

- Motivations: We need to **estimate user correlations on the fly**, during online recommendation.
  - **Clustering of <u>Linear</u> Bandits** [1, 2, 3, 4, 5, 6]**:**
    - Under **linear** stochastic contextual bandit settings: $r = \langle \theta_u, x \rangle + \eta$.
    - User **correlation intensity** between u, u' is measured by $\left\| \theta_u - \theta_{u'} \right\|_2$.
    - Adopt *combination of linear estimators* for reward estimation & exploration.
  - **Clustering of <u>Neural</u> Bandits** [7]**:**
    - Under **neural** stochastic contextual bandit settings: $r = h_u(x) + \eta$.
    - User clusters with identical preferences ($\forall u, u' \in \mathcal{N}, x \in \mathbb{R}^d$: $h_u(x) = h_{u'}(x)$).
    - Utilizing *gradient-based Meta-Learning* for reward estimation & exploration.

[1] Gentile et. al., Online clustering of bandits. ICML 2014.
[2] Li et. al., Improved algorithm on online clustering of bandits. IJCAI 2019.
[3] Nguyen et. al., Dynamic clustering of contextual multi-armed bandits. CIKM 2014.
[4] Gentile et. al., On context-dependent clustering of bandits. ICML 2017.
[5] Ban et. al., Local clustering in contextual multi-armed bandits. WWW 2021.
[6] Li et. al., Collaborative filtering bandits. SIGIR 2016.
[7] Ban et. al., Meta clustering of neural bandits. In submission.

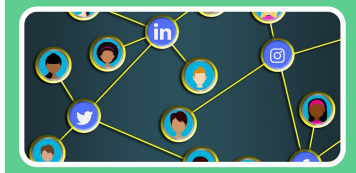
Roadmap

Introduction

- Background & Motivations
- Challenges

Online Clustering of Bandits

- Clustering of Linear Bandits
- Clustering of Neural Bandits



Graph Bandit Learning with Collaboration

- User side: Graph Neural Bandits
- Arm side: Neural Bandit with Arm Group Graph
- Other Scenarios: Bandit Learning with Graph Feedback & Online Graph Classification with Neural Bandit

Application in Recommender Systems

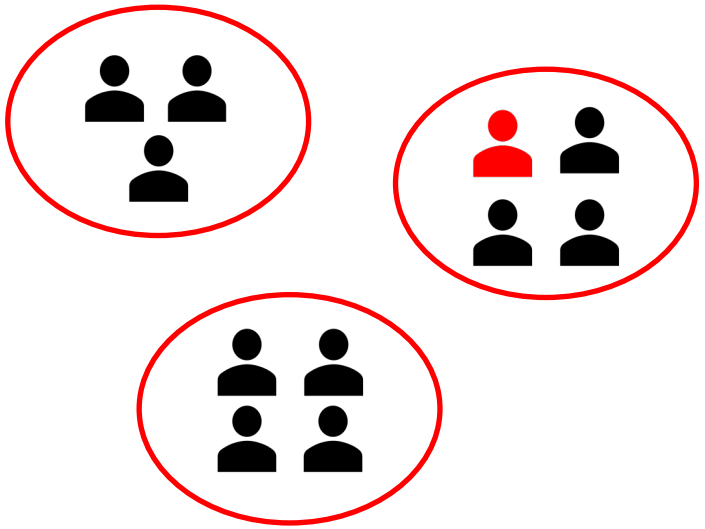
- Multi-facet Personalized Recommendation

Collaborative Exploration: Graph Bandits Learning

User (Bandit)

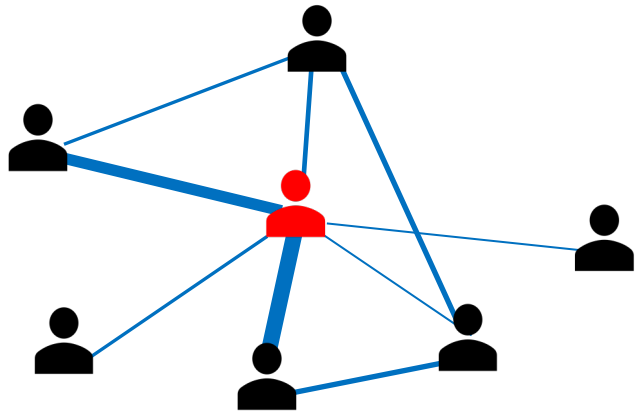
Clustering of Bandits [1,2]

- **Coarse-grained** user correlations:
  - Users within the same cluster share **identical preferences**.
  - **Contribute equally** to serving user.



Graph Bandits Learning [3]

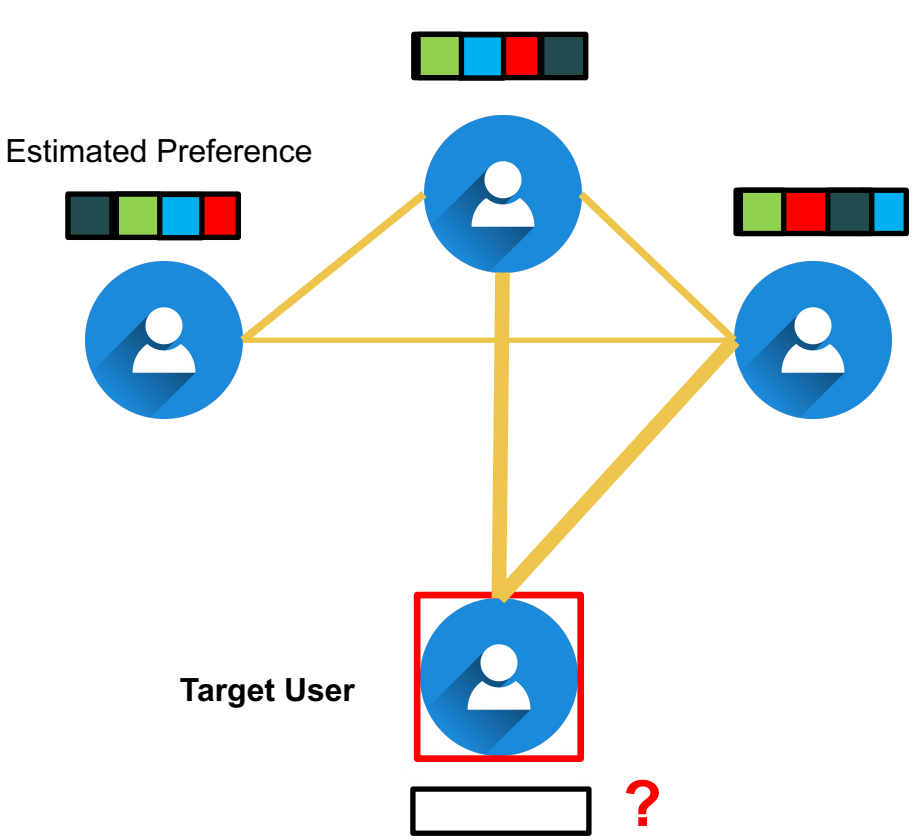
- **Fine-grained** user correlations:
  - Heterogeneity of users is preserved.
  - **Contribute differently** to serving user.



User correlation with strength (**Unknown**)

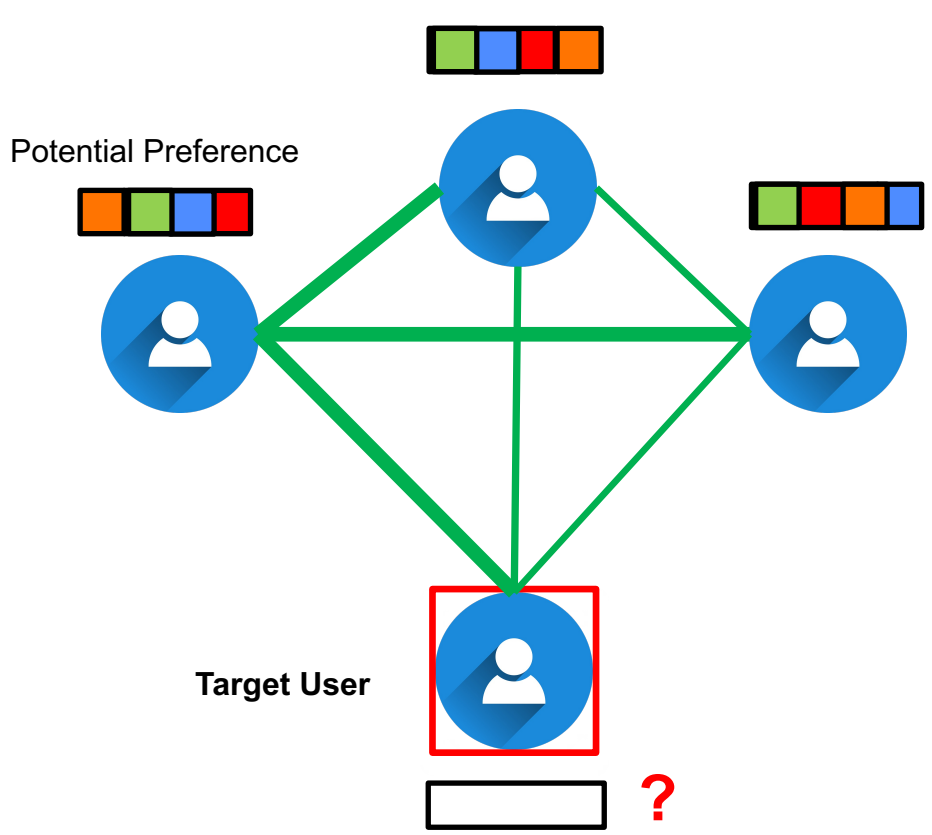
[1] Claudio Gentile, et al. 2014. Online clustering of bandits. In ICML. 757–765.
[2] Shuai Li, et al. 2019. Improved Algorithm on Online Clustering of Bandits. In IJCAI. 2923–2929.
[3] Y. Qi, Y. Ban*, and J. He. Graph neural bandits. KDD 2023.

GNB: Exploitation and Exploration Graphs



User Exploitation Graph

Correlation w.r.t. Exploitation



User Exploration Graph

Correlation w.r.t. Exploration

• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: Problem Definition

- **For each round** $t \in [T]$ **:**
  - Receive a target user $u_t \in \mathcal{U}$, and candidate arms (items) $\mathcal{X}_t$.
    - $\mathcal{X}_t = \{\boldsymbol{x}_{i,t} \in \mathbb{R}^d, \ (\text{e.g.,} \ \text{  }) \}_{i \in [a]}$
  - Reward $r_{i,t} = h\left(\mathcal{G}_{i,t}^{(1),*}, u_t, x_{i,t}\right) + \epsilon_{i,t}$.

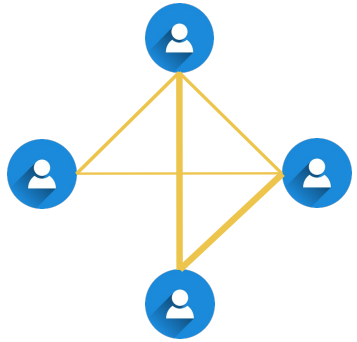
    Zero-mean noise
  - Learner **selects** arm $x_t \in \mathcal{X}_t$ as the recommendation.

• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: Problem Definition

- **For each round** $t \in [T]$ **:**
  - Receive a target user $u_t \in \mathcal{U}$, and candidate arms (items) $\mathcal{X}_t$.
    - $\mathcal{X}_t = \{\boldsymbol{x}_{i,t} \in \mathbb{R}^d, \ (\text{e.g.,}\ \text{  }\ )\}_{i\in[a]}$
  - Reward $r_{i,t} = h\left(\mathcal{G}_{i,t}^{(1),*}, u_t, x_{i,t}\right) + \epsilon_{i,t}$.
  - Learner **selects** arm $x_t \in \mathcal{X}_t$ as the recommendation.

- **Definition: User Correlation (Exploitation) Graph**
  - Given arm $\boldsymbol{x}_{i,t}$, **unknown** user exploitation graph $\mathcal{G}_{i,t}^{(1),*} = \left(\mathcal{U}, E, W_{i,t}^{(1),*}\right)$:
    - $\mathcal{U}$: set of nodes (**users**)
    - $E = \{e(u,u')\}_{u,u'\in\mathcal{U}}$: set of edges
    - $W_{i,t}^{(1),*}$: set of edge weights
      $W_{i,t}^{(1),*} = \boldsymbol{\Psi}^{(1)}\left(\mathbb{E}[\boldsymbol{r}_{i,t} \mid \boldsymbol{u_1}, \boldsymbol{x}_{i,t}], \mathbb{E}[\boldsymbol{r}_{i,t} \mid \boldsymbol{u_2}, \boldsymbol{x}_{i,t}]\right)$



User correlations w.r.t. the expected reward (**Exploitation Graph**)

• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: User Exploration Graph

➢ **Definition: User Exploration Graph**

❑ For arm $x_{i,t}$, **unknown** user exploration graph

$$\mathcal{G}_{i,t}^{(2),*} = (\mathcal{U}, E, W_{i,t}^{(2),*})$$

Set of edge weights

❑ For users $u_1, u_2 \in \mathcal{U}$, the corresponding edge weight:

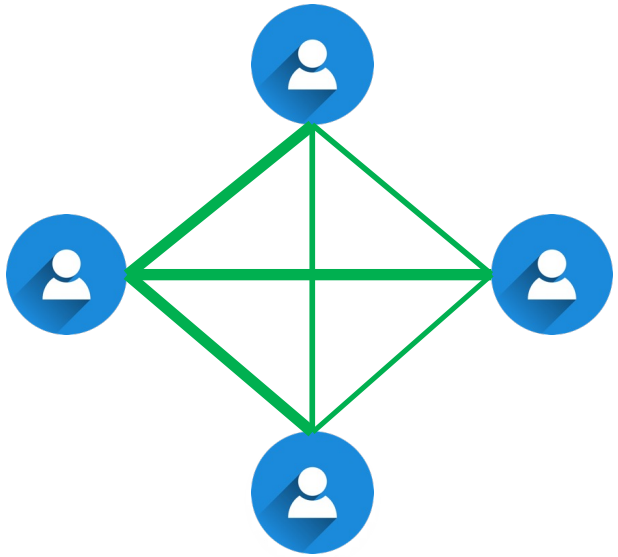
- $w_{i,t}^{(2),*}(u_1, u_2) = \Psi^{(2)}\left(\mathbb{E}[r_{i,t} | u_1, x_{i,t}] - f_{u_1}^{(1)}(x_{i,t}),\ \mathbb{E}[r_{i,t} | u_2, x_{i,t}] - f_{u_2}^{(1)}(x_{i,t})\right)$

Pre-defined mapping

Potential Gain

Potential Gain:

- $\mathbb{E}[r \mid u, x] - f_u^{(1)}(x)$
- Measures the uncertainty for the reward estimation



User correlations w.r.t. the Potential Gain (**Exploration Graph**)

- Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.
- Yikun Ban, et al. EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits. In ICLR 2022.

GNB: Problem Definition

- **For each round** $t \in [T]$ **:**
  - Receive a target user $u_t \in \mathcal{U}$, and candidate arms $\mathcal{X}_t$.
    - $\mathcal{X}_t = \{\boldsymbol{x}_{i,t} \in \mathbb{R}^d, \ \text{(e.g., )}\}_{i\in[a]}$
  - Reward $r_{i,t} = h\left(\mathcal{G}_{i,t}^{(1),*}, u_t, x_{i,t}\right) + \epsilon_{i,t}$.
  - Learner **selects** arm $x_t \in \mathcal{X}_t$ as the recommendation.

- **Definition: User Correlation (Exploitation) Graph**
  - Given arm $\boldsymbol{x}_{i,t}$, **unknown** user exploitation graph $\mathcal{G}_{i,t}^{(1),*} = (\mathcal{U}, E, W_{i,t}^{(1),*})$
    - $W_{i,t}^{(1),*}$: set of edge weights
  - For users $u_1, u_2 \in \mathcal{U}$, the corresponding edge weight:
    - $w_{i,t}^{(1),*}(u_1, u_2) = \boldsymbol{\Psi}^{(1)}\left(\mathbb{E}[\boldsymbol{r}_{i,t} \mid \boldsymbol{u_1}, \boldsymbol{x}_{i,t}], \mathbb{E}[\boldsymbol{r}_{i,t} \mid \boldsymbol{u_2}, \boldsymbol{x}_{i,t}]\right)$

- **Objective: Minimizing Pseudo Regret**

Chosen arm reward

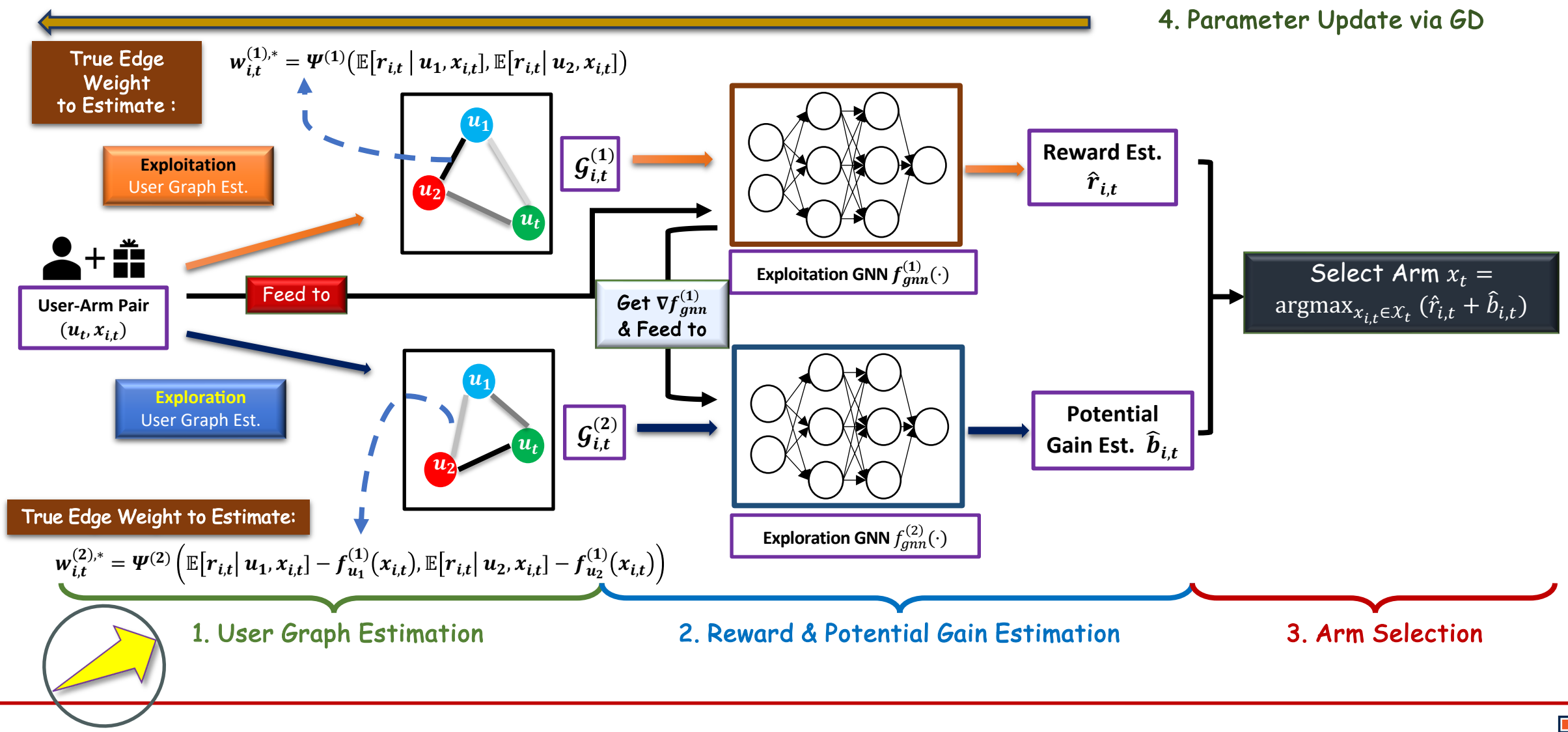
$$\mathrm{R(T)} = \sum_{t=1}^{T} \mathbb{E}[r_t^* - r_t]$$

Optimal arm reward

$$\mathbb{E}\,[r_t^*] = \max_{i\in[a]} \mathbb{E}[r_{i,t}]$$

• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

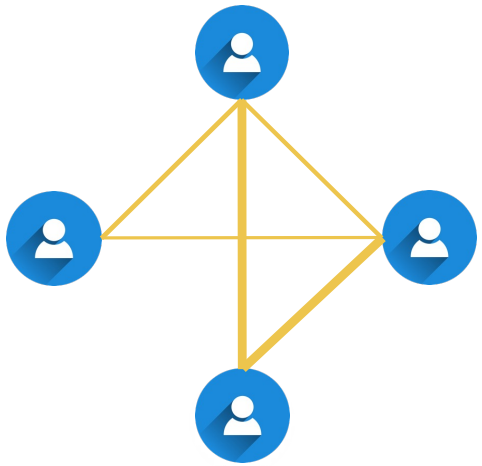
GNB: Framework Overview



• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

User Exploitation Graph Estimation

User Preference (expected reward) Estimation:
- Estimated by **user exploitation networks** $\left\{f_u^{(1)}\right\}_{u\in\mathcal{U}}$
- Approximating $\mathbb{E}[\boldsymbol{r} \mid \boldsymbol{u}, \boldsymbol{x}]$
- Input: $\boldsymbol{x}$ Label: $\boldsymbol{r}$



User correlations w.r.t. the expected reward (**Exploitation Graph**)

➢ **User Exploitation Graph Estimation:**

❑ Given arm $\boldsymbol{x}_{i,t}$, **estimated** user exploitation graph $\mathcal{G}_{i,t}^{(1)} = (\mathcal{U}, E, W_{i,t}^{(1)})$

- $W_{i,t}^{(1)}$: set of estimated edge weights

❑ For users $u_1, u_2 \in \mathcal{U}$, estimated edge weight
- $w_{i,t}^{(1)}(u_1, u_2) = \boldsymbol{\Psi}^{(1)}\left(f_{u_1}^{(1)}(\boldsymbol{x}_{i,t}),\ f_{u_2}^{(1)}(\boldsymbol{x}_{i,t})\right)$

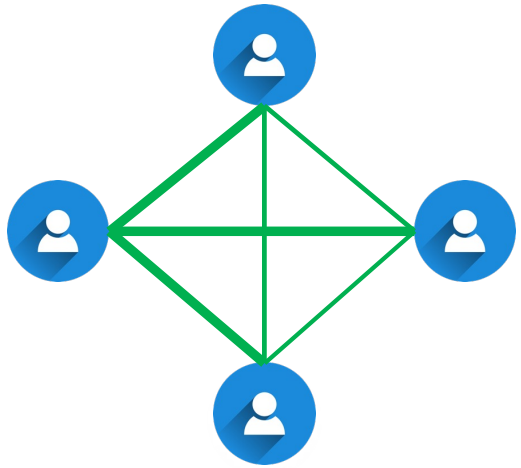
Estimated User Preference

• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

User Exploration Graph Estimation

Potential Gain:

- Estimated by user **exploration** networks $\left\{f_u^{(2)}\right\}_{u\in\mathcal{U}}$
  - **Input**: $\nabla f_u^{(1)}(\boldsymbol{x})$ -- the **gradients** of $f_u^{(1)}$.
  - **Label**: $\boldsymbol{r_u} - \boldsymbol{f_u^{(1)}}(\boldsymbol{x})$.



User correlations w.r.t. the Potential Gain (**Exploration Graph**)

➢ User Exploration Graph Estimation:

❑ Given arm $\boldsymbol{x}_{i,t}$, **estimated** user exploration graph $\mathcal{G}_{i,t}^{(2)} = (\mathcal{U}, E, W_{i,t}^{(2)})$

Edge weight estimations

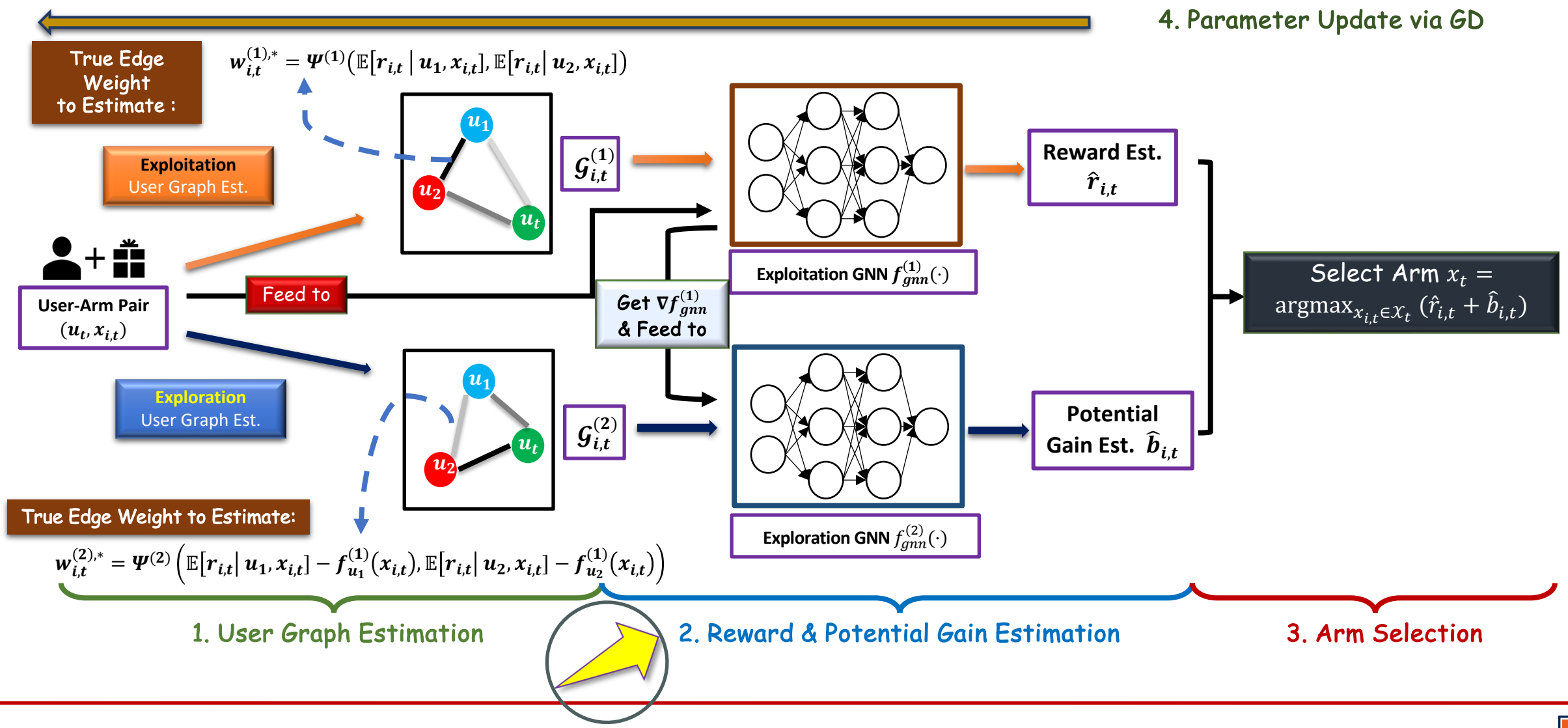
❑ For users $u_1, u_2 \in \mathcal{U}$, estimated edge weight

- $w_{i,t}^{(2)}(u_1, u_2) =$ $\boldsymbol{\Psi}^{(2)}\left(f_{u_1}^{(2)}(\nabla f_{u_1}^{(1)}(\boldsymbol{x}_{i,t})),\ f_{u_2}^{(2)}(\nabla f_{u_2}^{(1)}(\boldsymbol{x}_{i,t}))\right)$

Estimated Potential Gain

- Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.
- Yikun Ban, et al. EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits. In ICLR 2022.

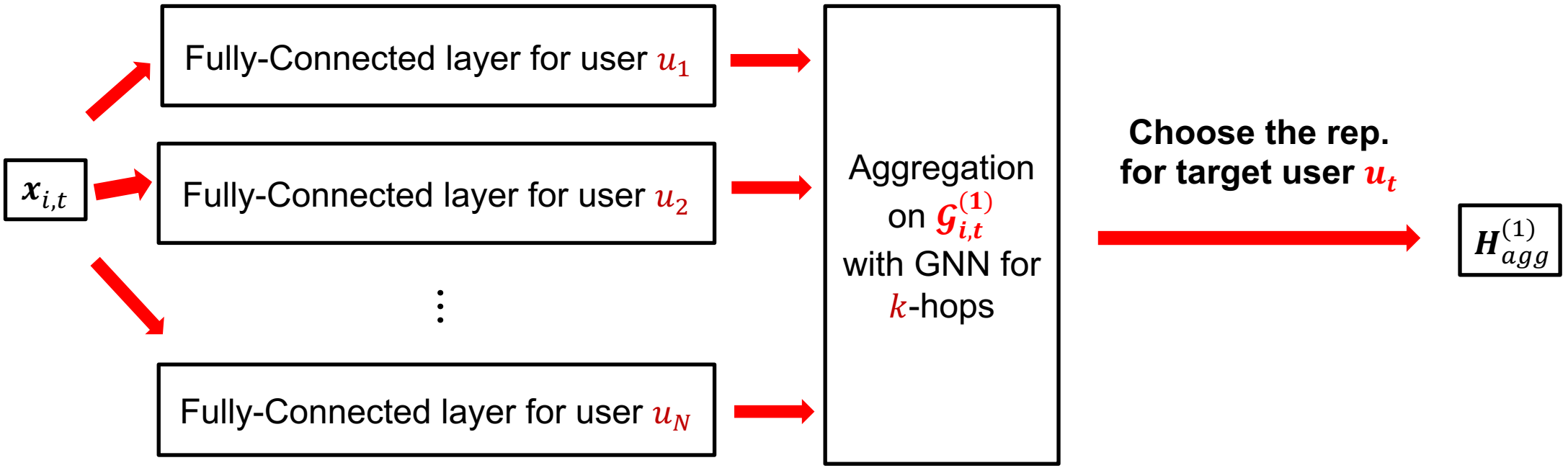
GNB: Framework Overview



- Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: Aggregation on User Exploitation Graph

- **For each arm $x_{i,t} \in \mathcal{X}_t$, reward estimation with estimated user exploitation graph $\mathcal{G}_{i,t}^{(1)}$.**
  - **Given target user u_t, obtain User-specific Arm Representation H_{agg}:**

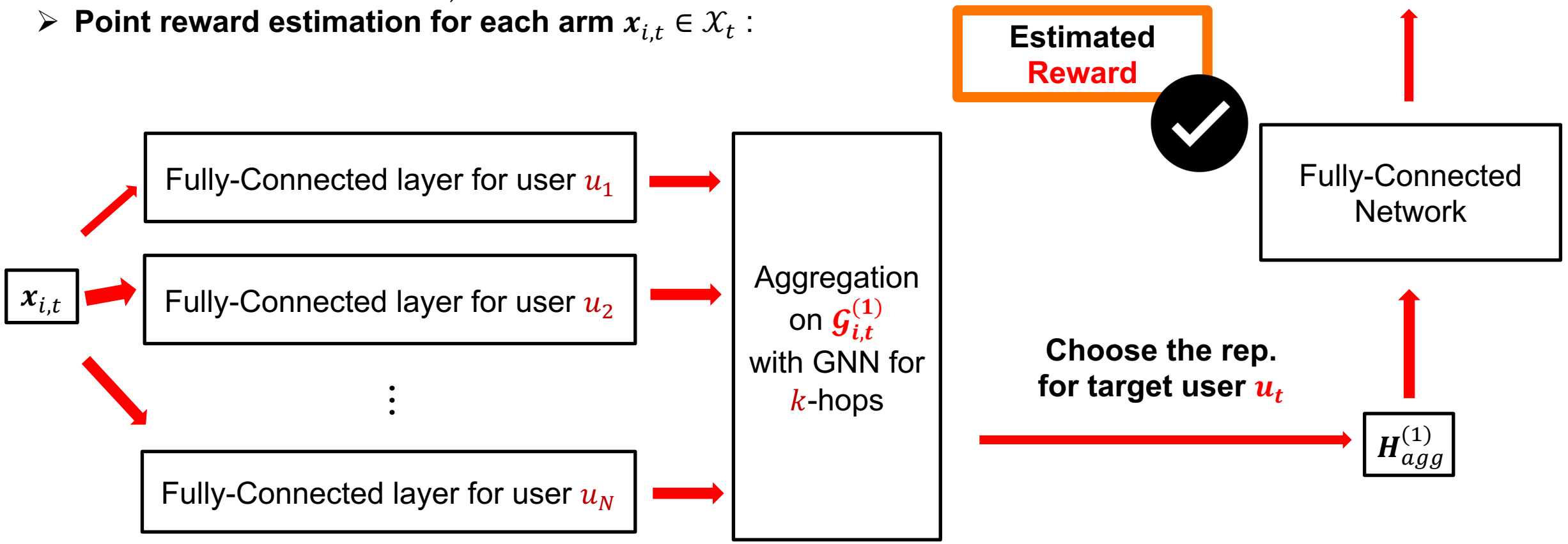


• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: Arm Reward Estimation

- **For each arm $x_{i,t} \in \mathcal{X}_t$, reward estimation with estimated user exploitation graph $\mathcal{G}_{i,t}^{(1)}$.**
  - **Point reward estimation for each arm $x_{i,t} \in \mathcal{X}_t$:**

$$\hat{r}_{i,t} = f_{gnn}^{(1)}\left(x_{i,t}, \mathcal{G}_{i,t}^{(1)}; \left[\Theta_{gnn}^{(1)}\right]_{t-1}\right)$$



• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: Potential Gain Estimation

- **For each arm $x_{i,t} \in \mathcal{X}_t$, reward estimation with estimated user exploration graph $\mathcal{G}_{i,t}^{(2)}$.**
  - **Potential gain estimation for each arm $x_{i,t} \in \mathcal{X}_t$:**

$$\hat{b}_{i,t} = f_{gnn}^{(2)}\left(\nabla_\Theta f_{gnn}^{(1)}, \mathcal{G}_{i,t}^{(2)}; \left[\boldsymbol{\Theta}_{gnn}^{(2)}\right]_{t-1}\right)$$

Estimated Potential Gain

Fully-Connected Network

$\boldsymbol{H}_{agg}^{(2)}$

Choose the rep. for target user u_t

Aggregation on $\mathcal{G}_{i,t}^{(2)}$ with GNN for k-hops

Fully-Connected layer for user u_1

Fully-Connected layer for user u_2

⋮

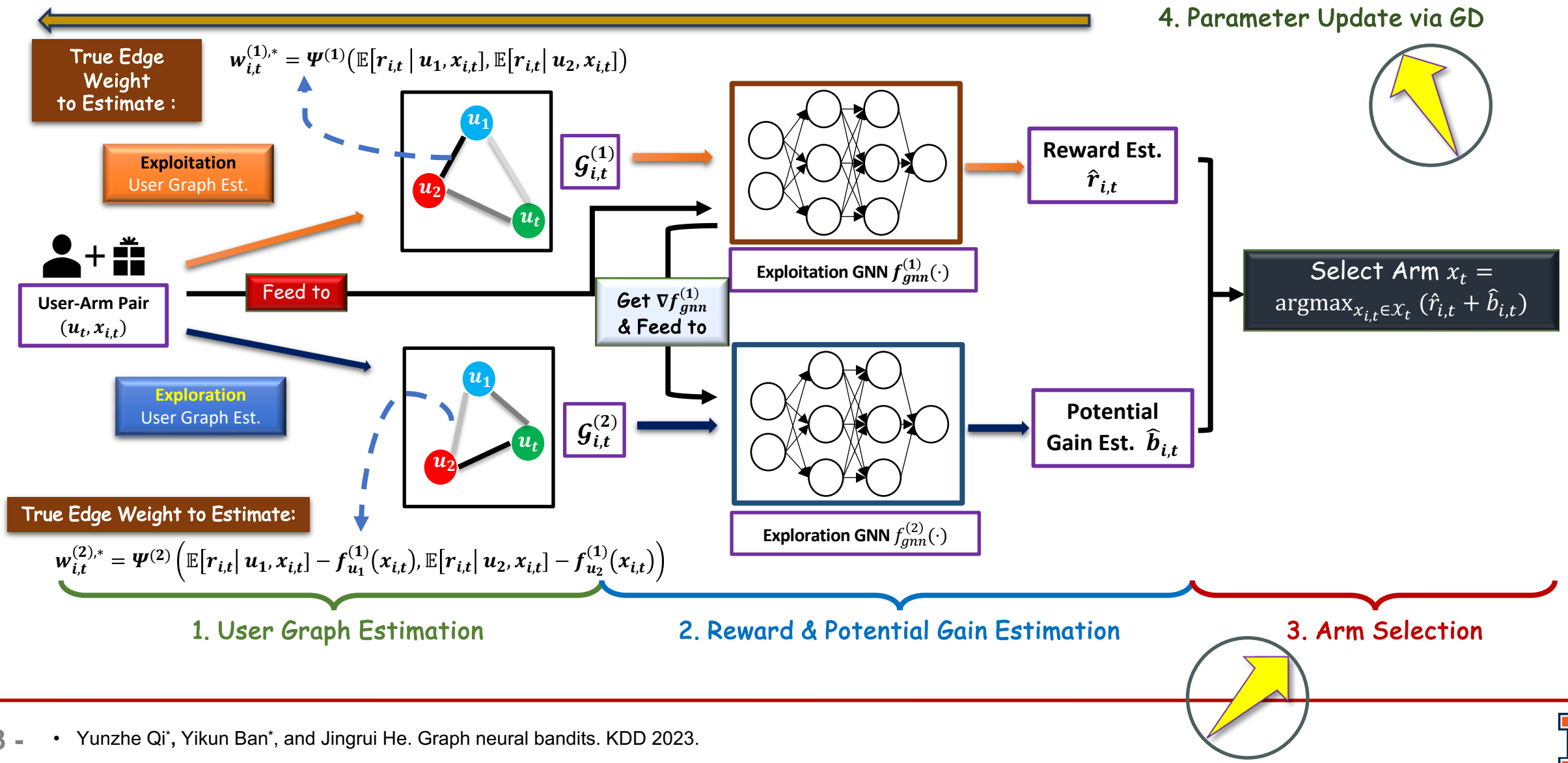
Fully-Connected layer for user u_N

$\nabla_\Theta f_{gnn}^{(1)}\left(\mathcal{G}_{i,t}^{(1)}, \boldsymbol{x}_{i,t}; \left[\boldsymbol{\Theta}_{gnn}^{(1)}\right]_{t-1}\right)$

Input: Gradients of $f_{gnn}^{(1)}$ in terms of $x_{i,t}$

- Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.
- Yikun Ban, et al. EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits. In ICLR 2022.

GNB: Framework Overview



• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: Arm Selection & Training

GNN + User Exploitation Graph

GNN + User Exploration Graph

- **Arm Selection Strategy:**
  - Select arm $\boldsymbol{x_t} = \text{argmax}_{x_{i,t} \in \mathcal{X}_t} (\hat{r}_{i,t} + \hat{b}_{i,t})$.

    Estimated Reward Estimated Potential Gain
  - Receive the corresponding true reward r_t.

- Train the model parameters (User models + GNN models) with **Gradient Descent (GD)**.
- Update user graphs.

• Yunzhe Qi*, Yikun Ban*, and Jingrui He. Graph neural bandits. KDD 2023.

GNB: Theoretical Analysis

- **Pseudo regret for T rounds:**

$$\boldsymbol{R(T)} = \sum_{t=1}^{T} \mathbb{E}[(r_t^* - r_t)]$$

- **Given sufficiently large network width m** (over-parameterization), under **mild assumptions**, with the probability at least $1 - \delta$**:**

$$R(T) \leq \sqrt{T} \cdot \left(\mathcal{O}(L\xi_L) \cdot \sqrt{2\log\left(\frac{Tn \cdot a}{\delta}\right)}\right) + \sqrt{T} \cdot \mathcal{O}(L) + \mathcal{O}(\xi_L) + \mathcal{O}(1).$$

where n is the number of users, a is the number of arms in each round, and T is the number of rounds.

Remarks:

- ❑ Achieves the regret bound of $\mathcal{O}(\sqrt{T\log(nT)})$.
  - Existing works with user clustering need $\boldsymbol{\mathcal{O}}(\sqrt{nT\log(T)})$ for user collaboration modeling.
- ❑ Free of the terms d
  - d (arm context dimension, common in linear bandit works)

- Shuai Li, et al. 2019. Improved Algorithm on Online Clustering of Bandits. In IJCAI. 2923–2929.
- Shuai Li, et al. 2016. Collaborative filtering bandits. In SIGIR. 539–548.
- Dongruo Zhou, et al. 2020. Neural contextual bandits with ucb-based exploration. In ICML. 11492–11502.
- Yikun Ban, et al. 2022. Neural Collaborative Filtering Bandits via Meta Learning. arXiv preprint arXiv:2201.13395 (2022).

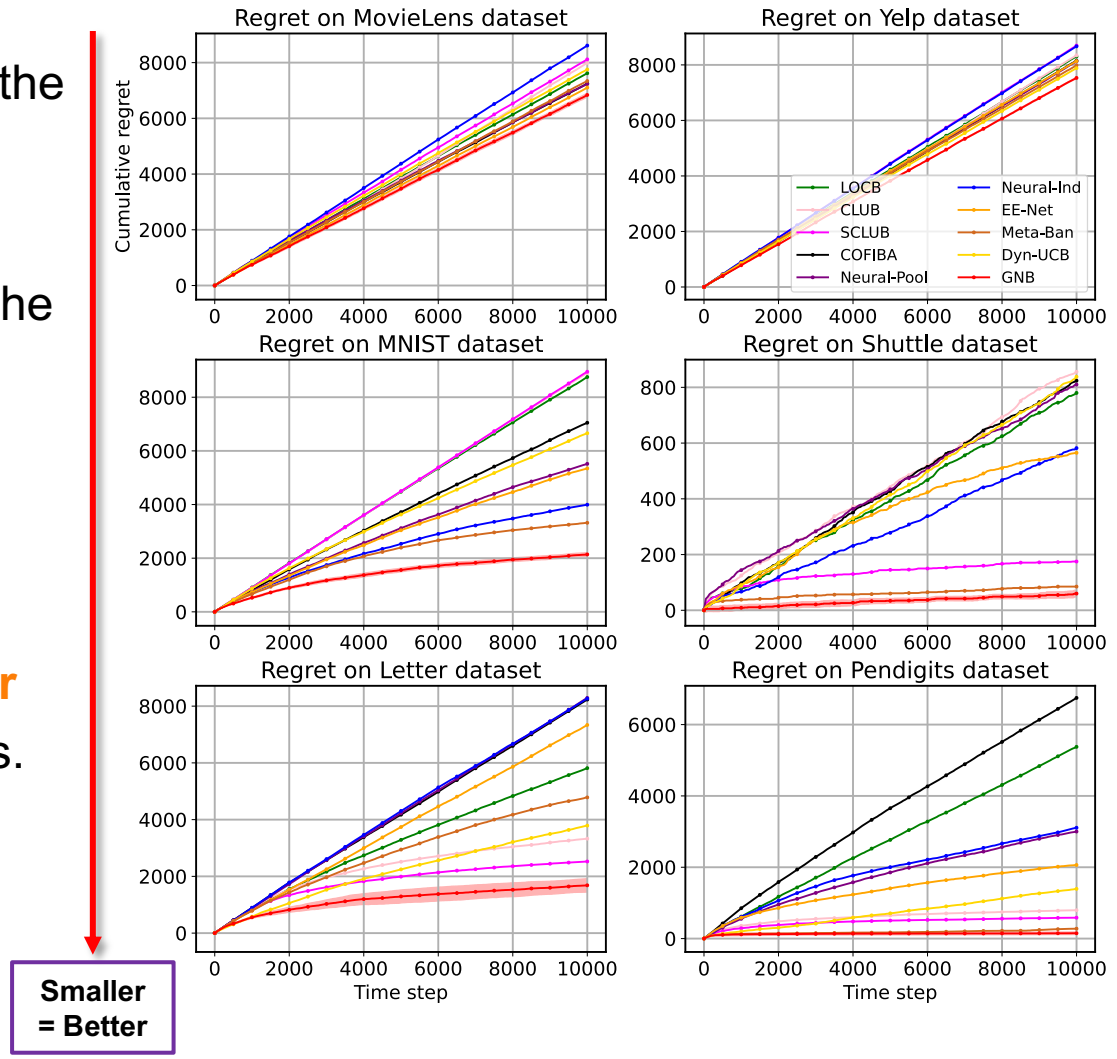
Experiments: Real Data Sets

➢ Experiment settings:

- ❑ Under **online recommendation** settings, we evaluate the proposed GNB framework on **six** real data sets with different specifications.
- ❑ We include **nine** state-of-the-art related algorithms as the baselines, including both linear and neural algorithms.

➢ Summary of experiment results:

- ❑ **Neural algorithms** generally perform better than **linear ones**, with the representation power of neural networks.
- ❑ GNB can generally achieve the **best performance** against the strong baselines.



- Shuai Li, et al. 2019. Improved Algorithm on Online Clustering of Bandits. In IJCAI. 2923–2929.
- Shuai Li, et al. 2016. Collaborative filtering bandits. In SIGIR. 539–548.
- Dongruo Zhou, et al. 2020. Neural contextual bandits with ucb-based exploration. In ICML. 11492–11502.
- Yikun Ban, et al. 2022. Neural Collaborative Filtering Bandits via Meta Learning. arXiv preprint arXiv:2201.13395 (2022).
- Trong T Nguyen and Hady W Lauw. 2014. Dynamic clustering of contextual multi-armed bandits. In CIKM. 1959–1962.
- Claudio Gentile, et al. 2014. Online clustering of bandits. In ICML. 757–765.
- Yikun Ban and Jingrui He. 2021. Local clustering in contextual multi-armed bandits. 2021. In WWW. 2335–2346.
- Yikun Ban, et al. 2022. EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits. In ICLR.

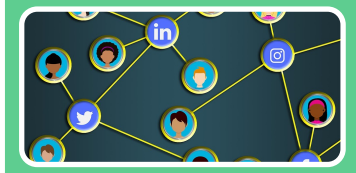
Roadmap

Introduction

- Background & Motivations
- Challenges

Online Clustering of Bandits

- Clustering of Linear Bandits
- Clustering of Neural Bandits



Graph Bandit Learning with Collaboration

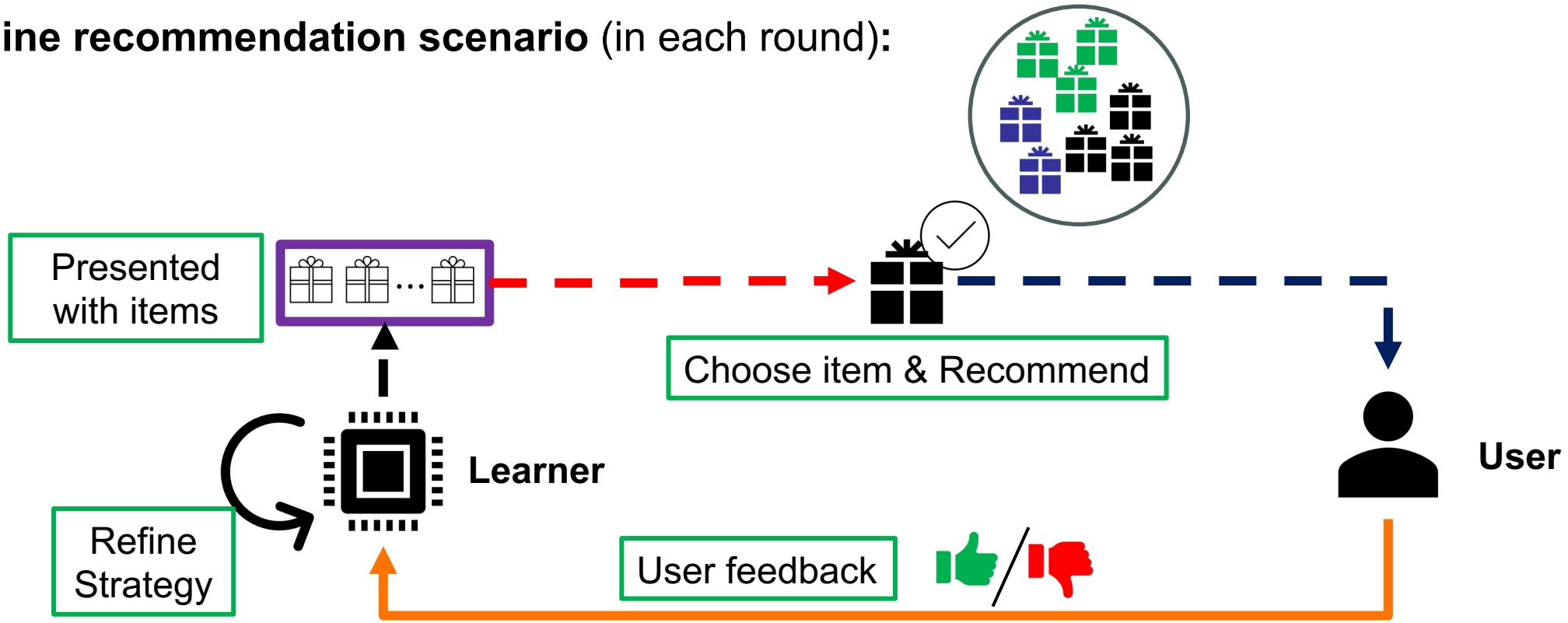
- User side: Graph Neural Bandits
- Arm side: Neural Bandit with Arm Group Graph
- Other Scenarios: Bandit Learning with Graph Feedback & Online Graph Classification with Neural Bandit

Bandits for Combo Recommendation

- Multi-facet Contextual Bandits

Online Recommendation with Arm Group Information

❑ **Online recommendation scenario** (in each round):



Leverage the available arm group information can help improve recommendation quality.

- Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

Arm Group Information

- **The group (category) information for arms (items) is commonly accessible:**
  - **Media contents:**
    - Music, Movies (grouped by genres)
  - **Text contents:**
    - Articles (grouped by literary styles)
  - **E-commerce:**
    - Restaurants (grouped by cuisine types)
  - Etc.

No existing **MAB** method trying to directly leverage the available arm group information.

• Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

Formal Problem Definition

- **Arm Groups:**
  - Assume a fixed pool $\mathcal{C}$ of $|\mathcal{C}| = N_c$ **arm groups**.
  - Each **arm group** $c \in \mathcal{C}$ (e.g., movie genre) relates to an arm distribution $\mathcal{D}_c$.
- **For each round** $t \in [T]$ **:**
  - Receive a set of arms $\mathcal{X}_t$, and the corresponding set of **arm groups** $\mathcal{C}_t \subseteq \mathcal{C}$.
    - $\mathcal{X}_t = \left\{ \boldsymbol{x}_{c,t}^{(i)} \in \mathbb{R}^{d_x}, \ \ (\text{e.g.,} \ \boxed{} \ ) \right\}_{c \in \mathcal{C}_t, \, i \in [n_{c,t}]}$
    - $\boldsymbol{x}_{c,t}^{(i)} \sim \mathcal{D}_c$
  - Reward $r_{c,t}^{(i)} = h\left(W^*, x_{c,t}^{(i)}\right) + \epsilon_{c,t}^{(i)}$.
    - Unknown affinity matrix for arm groups: $W^* \in \mathbb{R}^{N_c \times N_c}$
  - Learner **chooses** arm $x_t \in \mathcal{X}_t$.

- **Objective: Minimizing Pseudo Regret**

$$R(T) = \sum_{t=1}^{T} \mathbb{E}[(r_t^* - r_t)]$$

$$= \sum_{t=1}^{T} |h(W^*, x_t^*) - h(W^*, x_t)|$$

Optimal arm Chosen arm

• Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

Modeling with Arm Group Graph (AGG)

➢ **Apply Arm Group Graph (AGG) to model arm group correlations:**

➢ **In round** $t \in [T]$ **:**

❑ Undirected graph $\mathcal{G}_t = (V, E, W_t)$

- V: set of nodes
  - **Each node is an arm group** $c \in \mathcal{C}$**,** N_c nodes in total
- $E = \{e(c, c')\}_{c,c' \in \mathcal{C}}$: set of edges
- W_t: Set of edge weights

❑ Arm group correlations are modeled by the edge weights from set W_t.

➢ **True reward**: $r_{c,t}^{(i)} = h\left(W^*, x_{c,t}^{(i)}\right) + \epsilon_{c,t}^{(i)}$.

- Unknown affinity matrix: $W^* \in \mathbb{R}^{N_c \times N_c}$

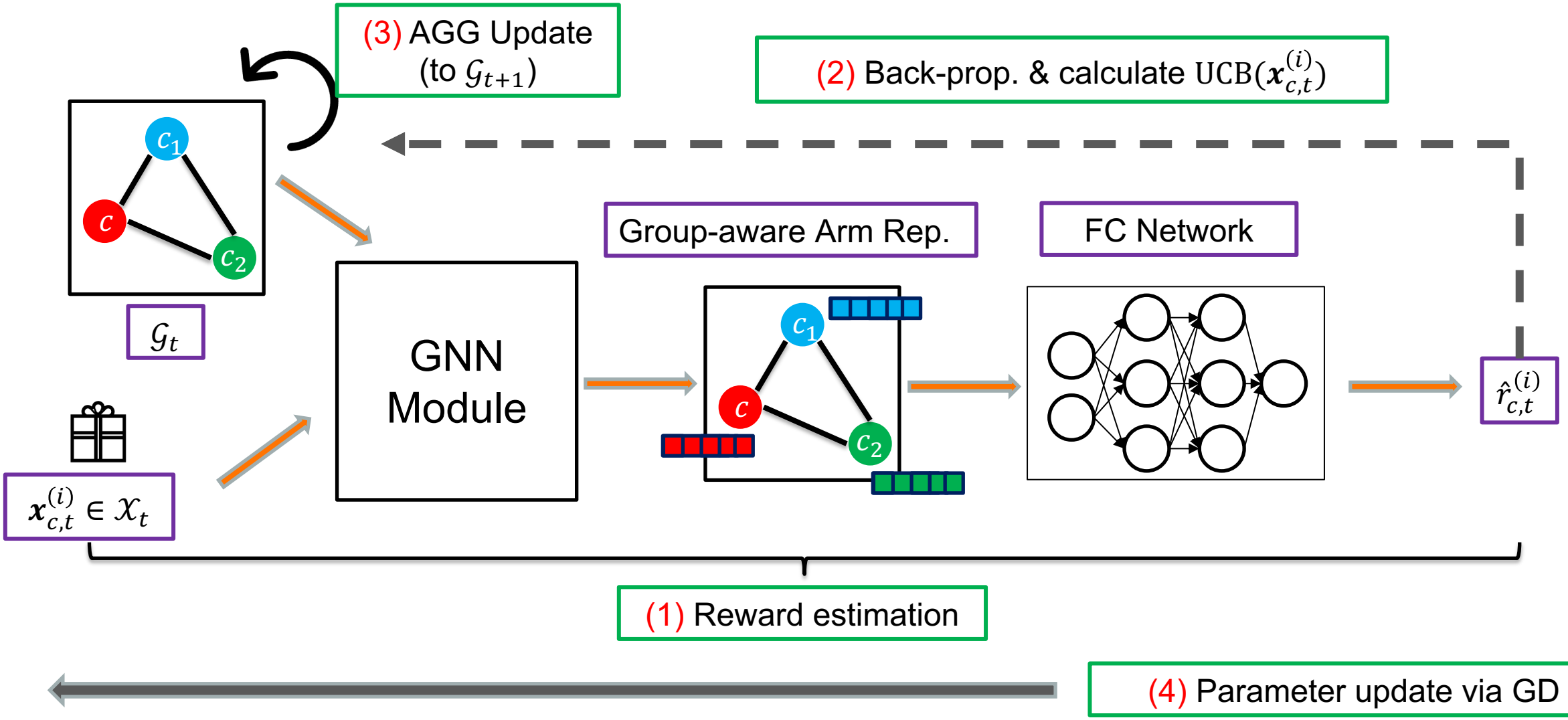
➢ **Unknown true graph:** $\mathcal{G}^*$

- Encoding the true arm group correlations.
- **In substitution for** W^***:**

$$r_{c,t}^{(i)} = h\left(\mathcal{G}^*, x_{c,t}^{(i)}\right) + \epsilon_{c,t}^{(i)}.$$

• Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

Proposed Framework: AGG-UCB



- Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

AGG-UCB: Arm Group Graph Estimation

- **Recall for Arm Groups:**
  - Assume a fix pool $\mathcal{C}$ of $|\mathcal{C}| = N_c$ arm groups.
  - Each group $c \in \mathcal{C}$ has a context distribution $\mathcal{D}_c$.

➢ **Definition: True edge weights**

❑ For $c,\ c' \in \mathcal{C}$, true edge weight in $\mathcal{G}^*$:

- $w^*(c, c') = \exp\left(\frac{-\left\|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_c}[\boldsymbol{\phi}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}'\sim\mathcal{D}_{c'}}[\boldsymbol{\phi}(\boldsymbol{x}')]\right\|^2}{\sigma_s}\right)$
- $\phi(\cdot)$: kernel mapping function

➢ **Arm Group Graph estimation:**

❑ Estimated edge weight in round $t \in [T]$:

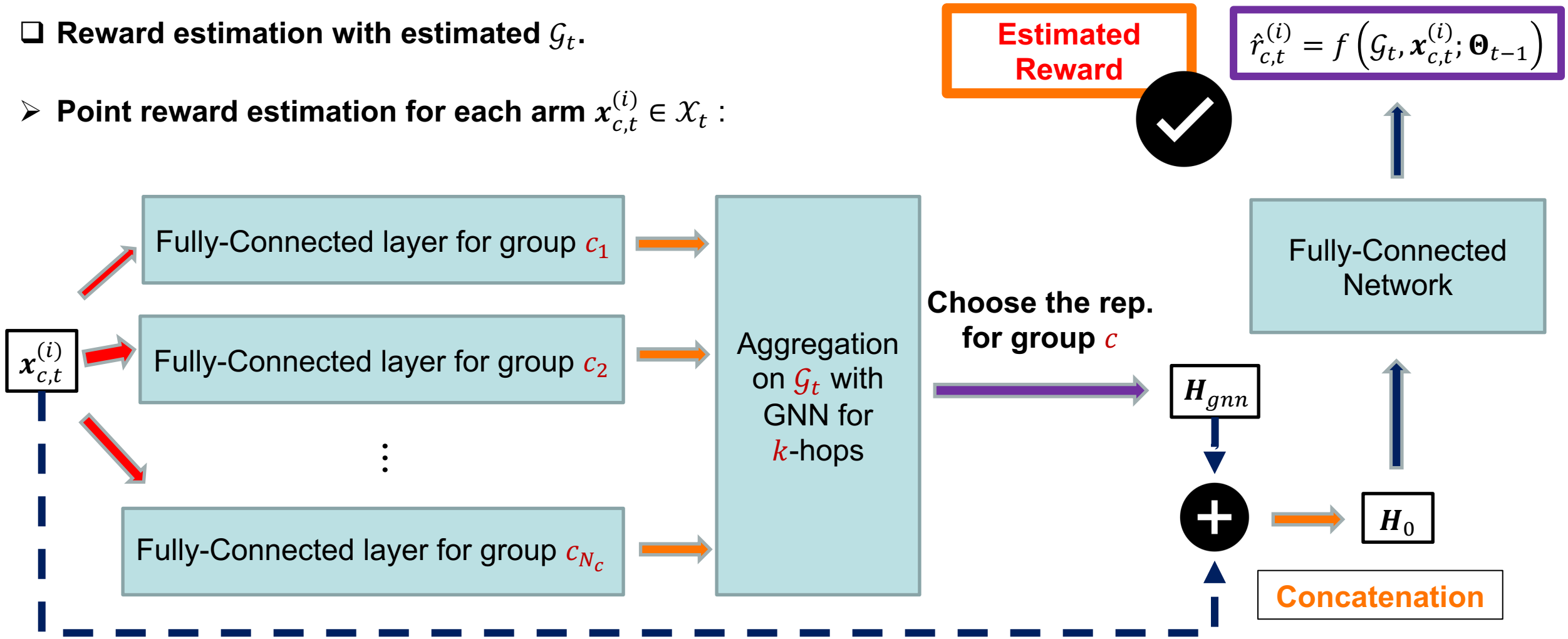
- $w_t(c, c') = \exp\left(\frac{-\|\Psi_t(\mathcal{D}_c) - \Psi_t(\mathcal{D}_{c'})\|^2}{\sigma_s}\right)$
- **Kernel Mean Embedding** [1]**:** $\Psi_t(\mathcal{D}_c)$

❑ $w_t(c, c') \in W_t$: weight for edge $e(c, c') \in E$ in graph $\mathcal{G}_t$

• Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. NeurIPS '11.

AGG-UCB: Arm Reward Estimation

- **Reward estimation with estimated $\mathcal{G}_t$.**
  - **Point reward estimation for each arm $\boldsymbol{x}_{c,t}^{(i)} \in \mathcal{X}_t$:**



- Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

AGG-UCB: Arm Selection & Training

- **Exploration with Upper Confidence Bound (UCB):**
  - The $\mathbf{UCB}(\cdot)$ satisfies :

$$\mathbb{P}\left(\left| \underbrace{f\left(\mathcal{G}_t, \boldsymbol{x}_{c,t}^{(i)}; \boldsymbol{\Theta}_{t-1}\right)}_{\text{Reward Est.}} - \underbrace{h\left(\mathcal{G}^*, \boldsymbol{x}_{c,t}^{(i)}\right)}_{\text{Exp. Reward}} \right| > \mathbf{UCB}\left(\boldsymbol{x}_{c,t}^{(i)}\right) \right) \leq \delta$$

- **Arm Selection Strategy:**
  - Select arm $\boldsymbol{x}_t = \operatorname{argmax}_{x_{c,t}^{(i)} \in \mathcal{X}_t} \left(\hat{r}_{c,t}^{(i)} + \gamma \cdot \mathbf{UCB}\left(\boldsymbol{x}_{c,t}^{(i)}\right) \right)$
  - Receive the corresponding true reward r_t

• Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

Theoretical and Empirical Results

- **Theoretical**: Given sufficiently large network width m, with the probability at least $1-\delta$:

$$R(T) \leq 2 \cdot (2B_4\sqrt{T} + 2 - B_4) + 2\sqrt{2\tilde{d}T\log(1+T/\lambda) + 2T} \cdot \left(\sqrt{\lambda}S + \sqrt{1 - 2\log(\delta/2) + (\tilde{d}\log(1+T/\lambda))}\right)$$

Achieves the regret bound of $\mathcal{O}(\tilde{d}\sqrt{T\log^2(T)\cdot\log(N_c)})$

- **Empirical**: Leveraging **arm group information with AGG-UCB** can improve good performances.

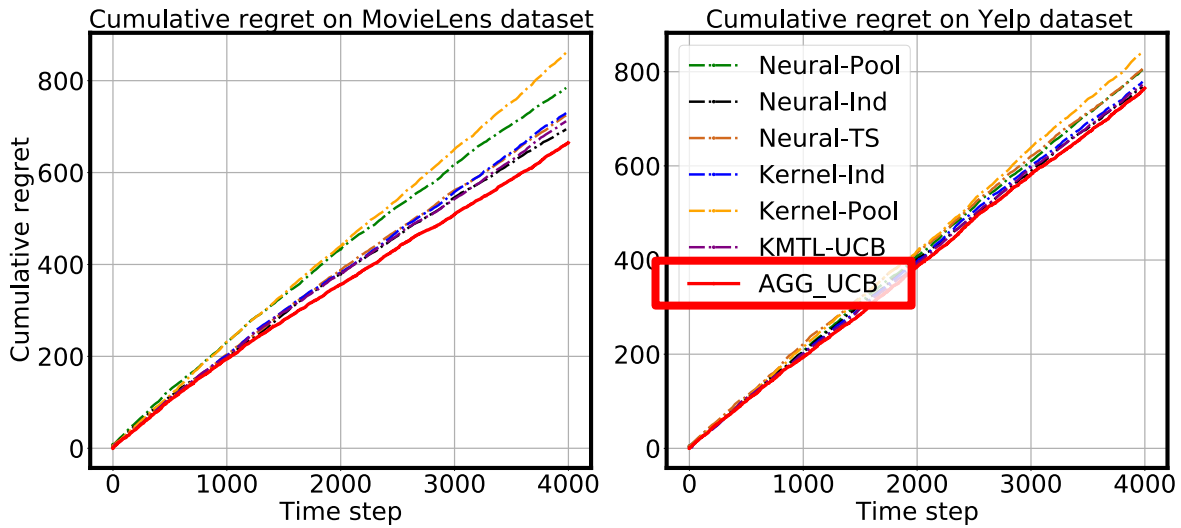


Figure 1: Cumulative regrets on recommendation data sets

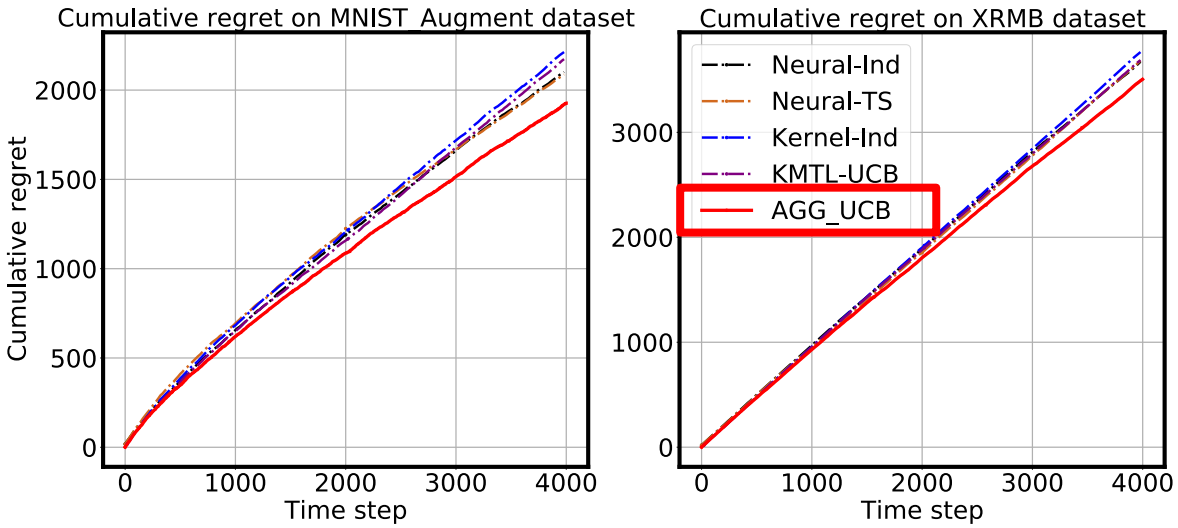


Figure 2: Cumulative regrets on Classification data sets

- Yunzhe Qi et. al., Neural bandit with arm group graph. KDD 2022.

Graph Bandit Learning: Other Scenarios

1. **Bandit Learning with Graph Feedback** [1]:

   - Arms are nodes on a graph $G = (V, E)$. In each round $t \in [T]$, the learner chooses one node $I_t \in V$.
   - Learner observes **reward for chosen arm** I_t, and **neighbor rewards** (e.g., out-neighbors in a directed graph).
   - **Objective**: minimizing the cumulative pseudo regret over T rounds.

2. **Optimal Graph Search with Bandit** [2]:

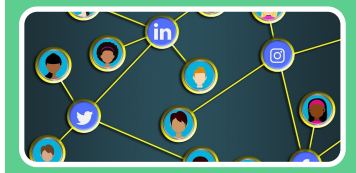
   - In each round $t \in [T]$, the learner aims to choose **one graph** $G_t \in \mathcal{G}$, from a **fixed** graph domain $\mathcal{G}$. Reward generated by $r_t = h(G_t) + \epsilon_t$.
   - **Objective**: minimizing the cumulative pseudo regret over T rounds.
   - **Application example:** material designing, drug search.

[1] Kong et.al., Simultaneously Learning Stochastic and Adversarial Bandits with General Graph Feedback. ICML 2022.
[2] Kassraie et al., Graph Neural Network Bandits. NeurIPS 2022.

Roadmap

Introduction
- Background & Motivations
- Challenges

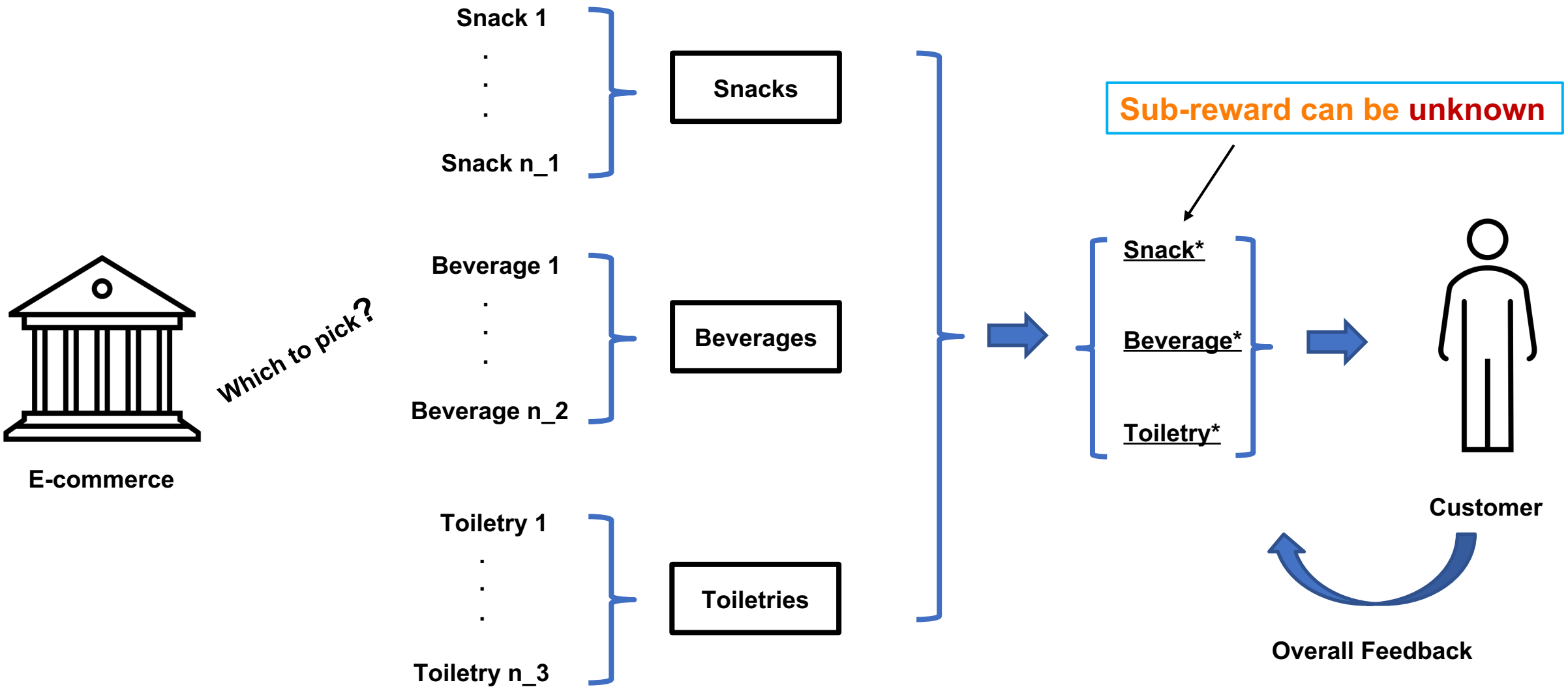
Online Clustering of Bandits
- Clustering of Linear Bandits
- Clustering of Neural Bandits



Graph Bandit Learning with Collaboration
- User side: Graph Neural Bandits
- Arm side: Neural Bandit with Arm Group Graph
- Other Scenarios: Bandit Learning with Graph Feedback & Online Graph Classification with Neural Bandit

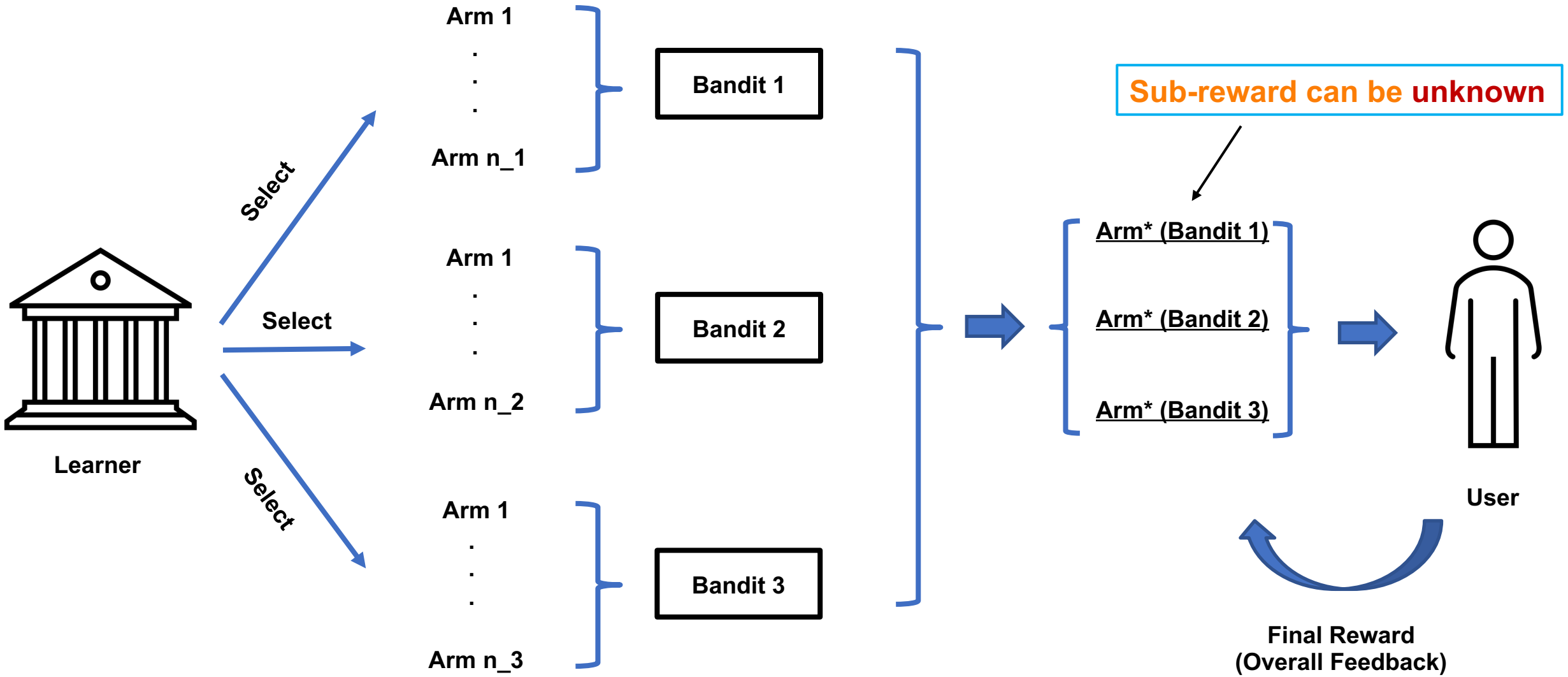
Bandits for Combo Recommendation
- Multi-facet Contextual Bandits

Motivated Case: Promotion Campaign



- Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

Application: Multi-facet Recommendation with Neural Bandits



- Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

Formal Definition of Multi-facet Contextual Bandit : In Round *t*

- **Sub-reward Functions (unknown):**

$$r_t^1 = h_1(x_t^1) \quad \text{(Linear or Non-linear)}$$

$$r_t^2 = h_2(x_t^2)$$

$$\vdots$$

$$r_t^K = h_K(x_t^K)$$

Assumtion1: $h_k(0) = 0, \ \forall\, k$

- **Final Reward Function (unknown):**

$$R_t = H\left(r_t^1, r_t^2, \dots, r_t^K\right) + \epsilon_t$$ ← Noise

Expectation: $H(X_t) = E[R_t|X_t] = H\left(r_t^1, r_t^2, \dots, r_t^K\right)$

Assumtion2: ***H* is $\bar{C}$- *Lipschitz continuous.***

- **Evaluation Measure: Regret**

$$Reg = E\left[\sum_t (R_t^* - R_t)\right]$$

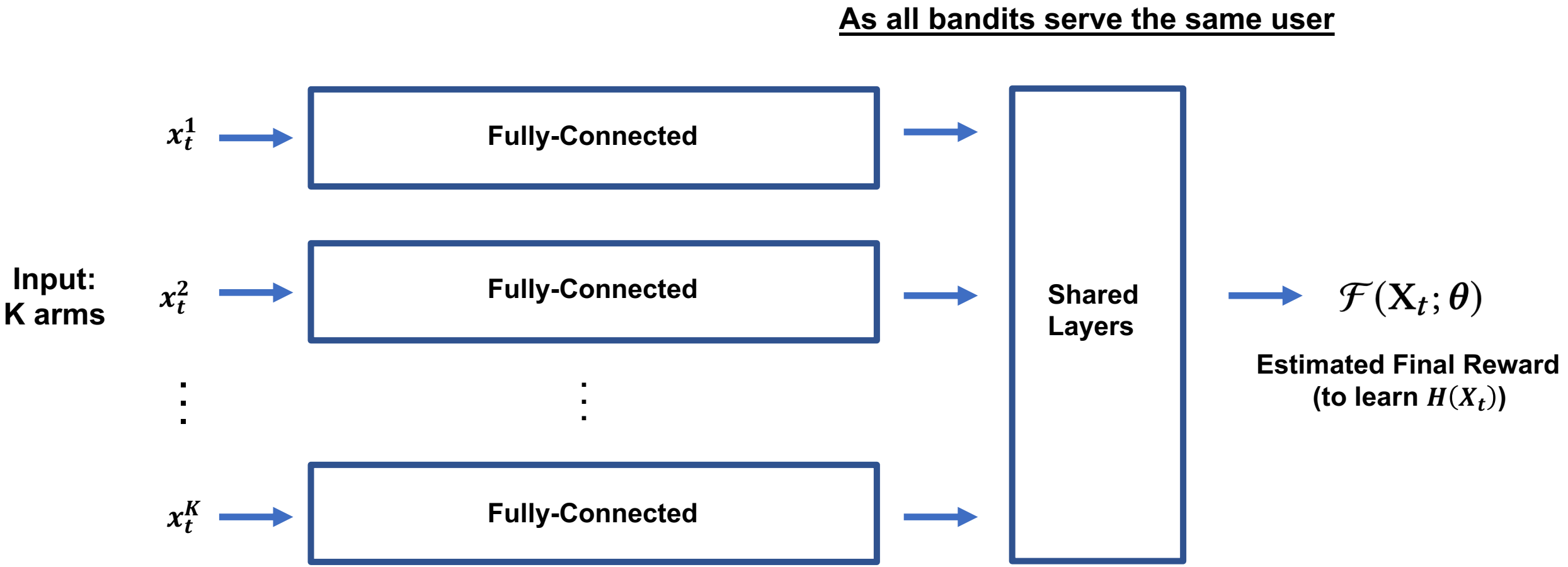
$$= \sum_t [H(X_t^*) - H(X_t)]$$

Optimal Final Reward **Received Final Reward**

- **Goal: Minimize the regret of T rounds.**

• Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

MuFasa: Exploitation (Neural Network Model)



- Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

MuFasa: Exploration (Upper Confidence Bound)

- **UCB:** $$\mathbb{P}\left(|\mathcal{F}(\mathbf{X}_t;\boldsymbol{\theta}_t) - \mathcal{H}(\mathbf{X}_t)| > \mathrm{UCB}(\mathbf{X}_t)\right) \leq \delta,$$

- **K selected arms are determined by:**

$$\mathbf{X}_t = \arg\max_{\mathbf{X}'_t \in \mathbf{S}_t} \left(\mathcal{F}(\mathbf{X}';\boldsymbol{\theta}_t) + \mathrm{UCB}(\mathbf{X}'_t)\right).$$

Where

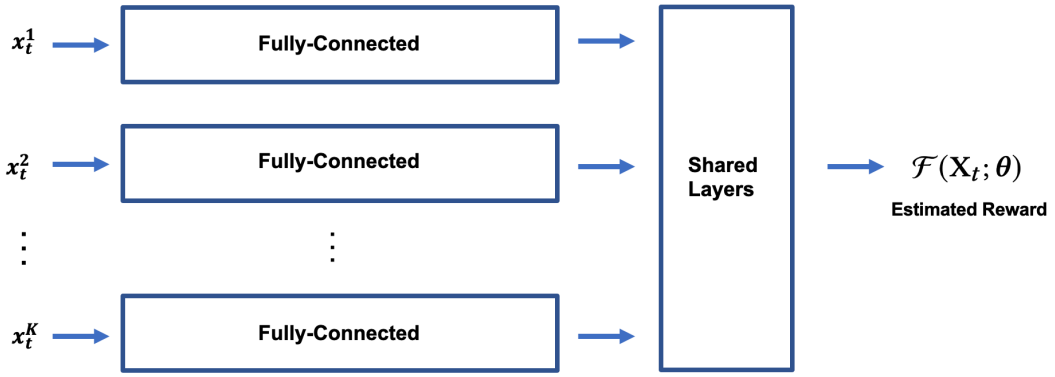
$$\mathbf{S}_t = \{(\mathbf{x}_t^1, \ldots, \mathbf{x}_t^k, \ldots, \mathbf{x}_t^K) \mid \mathbf{x}_t^k \in \mathbf{X}_t^k, k \in [K]\},$$

(all possible combinations of K arms)

• Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

MuFasa: Novel Upper Confidence Bound (UCB)

- **With the assembled neural framework (MuFasa):**



- **With probability at least $1-\delta$, the UCB holds**

$$|\mathcal{F}(\mathbf{X}_t;\boldsymbol{\theta}_t) - \mathcal{H}(\mathbf{X}_t)| \leq \bar{C}\sum_{k=1}^{K}\mathcal{B}^k + \mathcal{B}^F = UCB(\mathbf{X}_t),\ \textit{where}$$

$$\mathcal{B}^k = \gamma_1 \|g_k(\mathbf{x}_t^k;\boldsymbol{\theta}_t^k)/\sqrt{m_1}\|_{\mathbf{A}_t^{k-1}} + \gamma_2\left(\frac{\delta}{k+1}\right)\|g_k(\mathbf{x}_t^k;\boldsymbol{\theta}_0^k)/\sqrt{m_1}\|_{\mathbf{A}_t^{k\prime-1}} + \gamma_1\gamma_3 + \gamma_4$$

Error of facet-specific networks

$$\mathcal{B}^F = \gamma_1 \|G(\mathbf{f}_t;\boldsymbol{\theta}_t^{\Sigma})/\sqrt{m_2}\|_{\mathbf{A}_t^{F-1}} + \gamma_2\left(\frac{\delta}{k+1}\right)\|G(\mathbf{f}_t;\boldsymbol{\theta}_0^{\Sigma})/\sqrt{m_2}\|_{\mathbf{A}_t^{F\prime-1}} + \gamma_1\gamma_3 + \gamma_4$$

Error of shared network

- Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

Regret Analysis

$$\boldsymbol{Reg} = \boldsymbol{E}\left[\sum_{t}(\boldsymbol{R}_t^* - \boldsymbol{R}_t)\right]$$
$$= \sum_{t}[\boldsymbol{H}(\boldsymbol{X}_t^*) - \boldsymbol{H}(\boldsymbol{X}_t)]$$

- After T rounds, with probability at least $1-\delta$,

$$\boldsymbol{Reg} \leq (\bar{C}K+1)\sqrt{T}2\sqrt{\widetilde{P}\log(1+T/\lambda)+1/\lambda+1}$$
$$\cdot\left(\sqrt{(\widetilde{P}-2)\log\left(\frac{(\lambda+T)(1+K)}{\lambda\delta}\right)+1/\lambda+\lambda^{1/2}S+2}\right)+2(\bar{C}K+1),$$

- Achieve near-optimal regret bound $\widetilde{O}\left((K+1)\sqrt{T}\right)$, same as a single linear bandit $\widetilde{O}(\sqrt{T})$

- Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

All Sub-rewards Available (Different Final Reward Function)

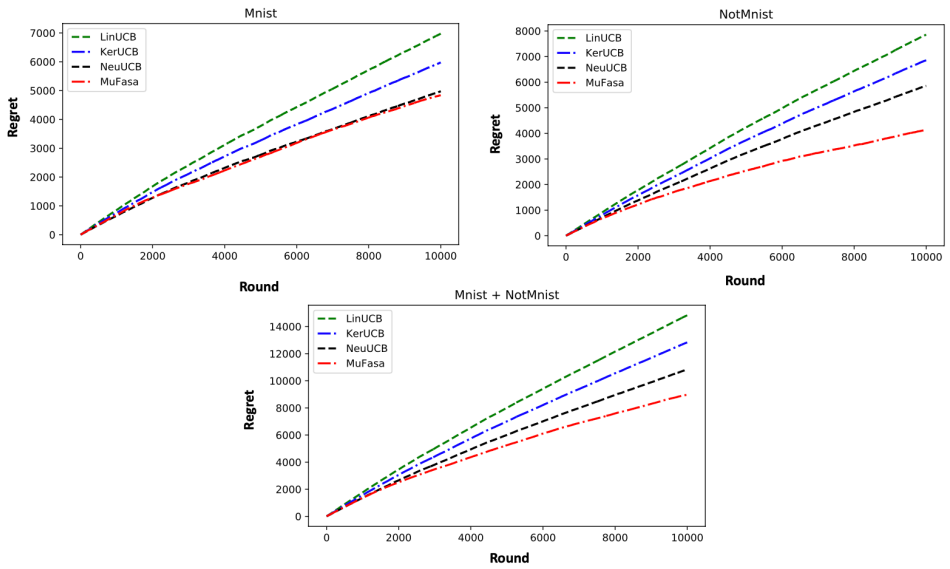


Figure: Regret comparison on Mnist+NotMnist with H_1.

$$H_1(\mathrm{vec}(\mathbf{r}_t)) = r_t^1 + r_t^2$$

Observation:

- **Superiority of MuFasa is slightly higher on H_2, compared to H_1.**

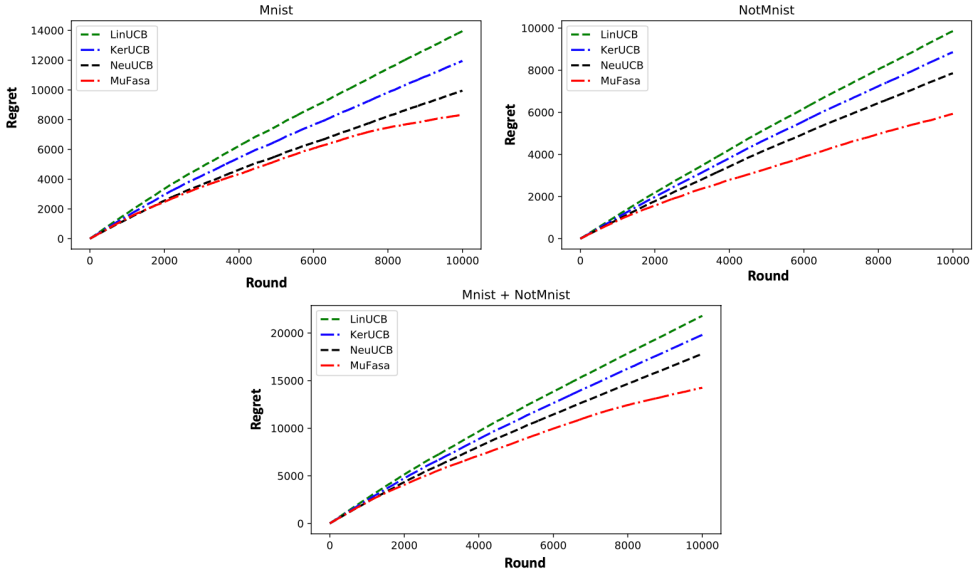


Figure: Regret comparison on Mnist+NotMnist with H_2.

$$H_2(\mathrm{vec}(\mathbf{r}_t)) = 2r_t^1 + r_t^2.$$

Insights:

- **MuFasa can select arms according to different weights of bandits (Bandit 1 has higher weight in H_2).**

- Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

Partial Sub-rewards Available

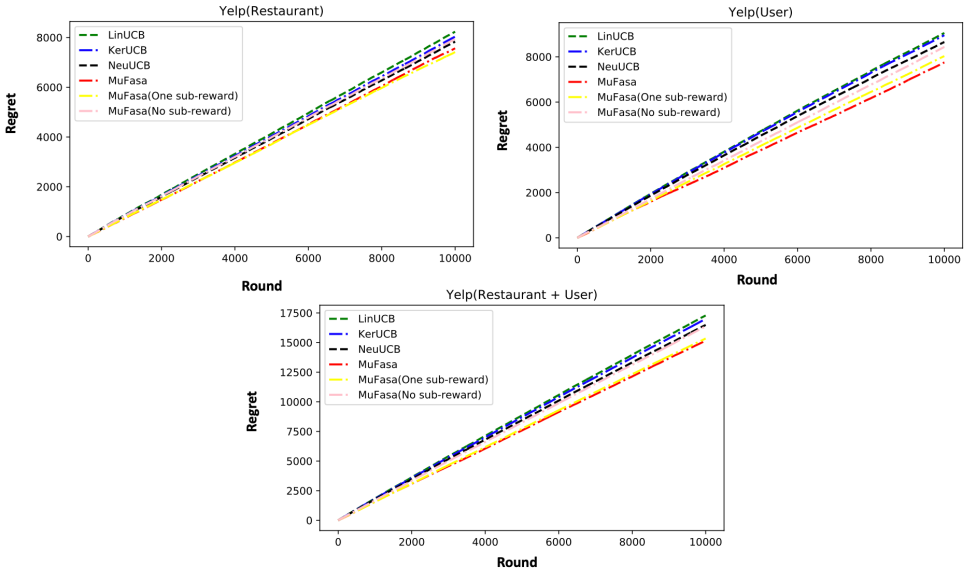


Figure: Regret comparison on Yelp with different reward availability.

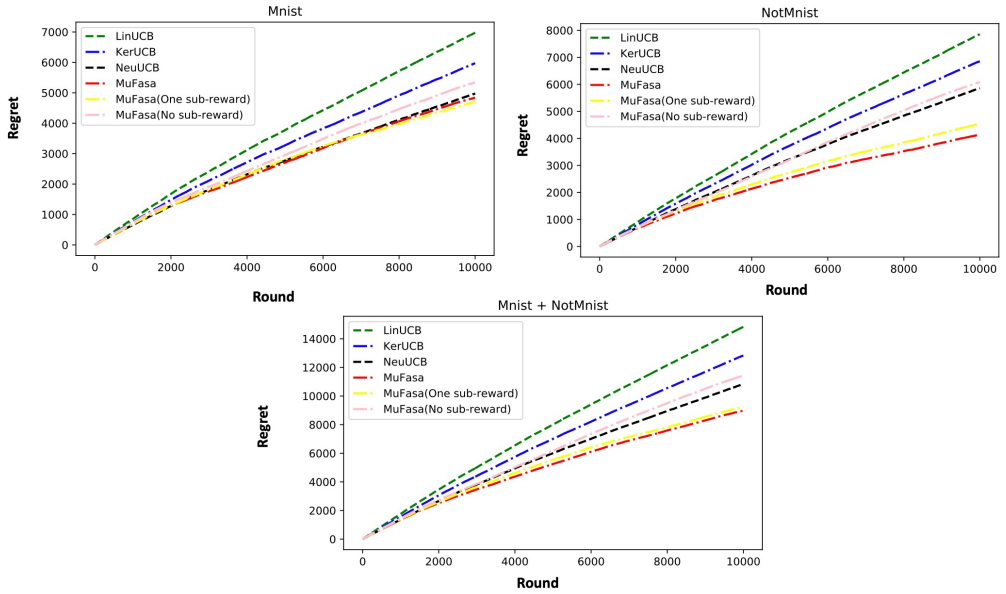


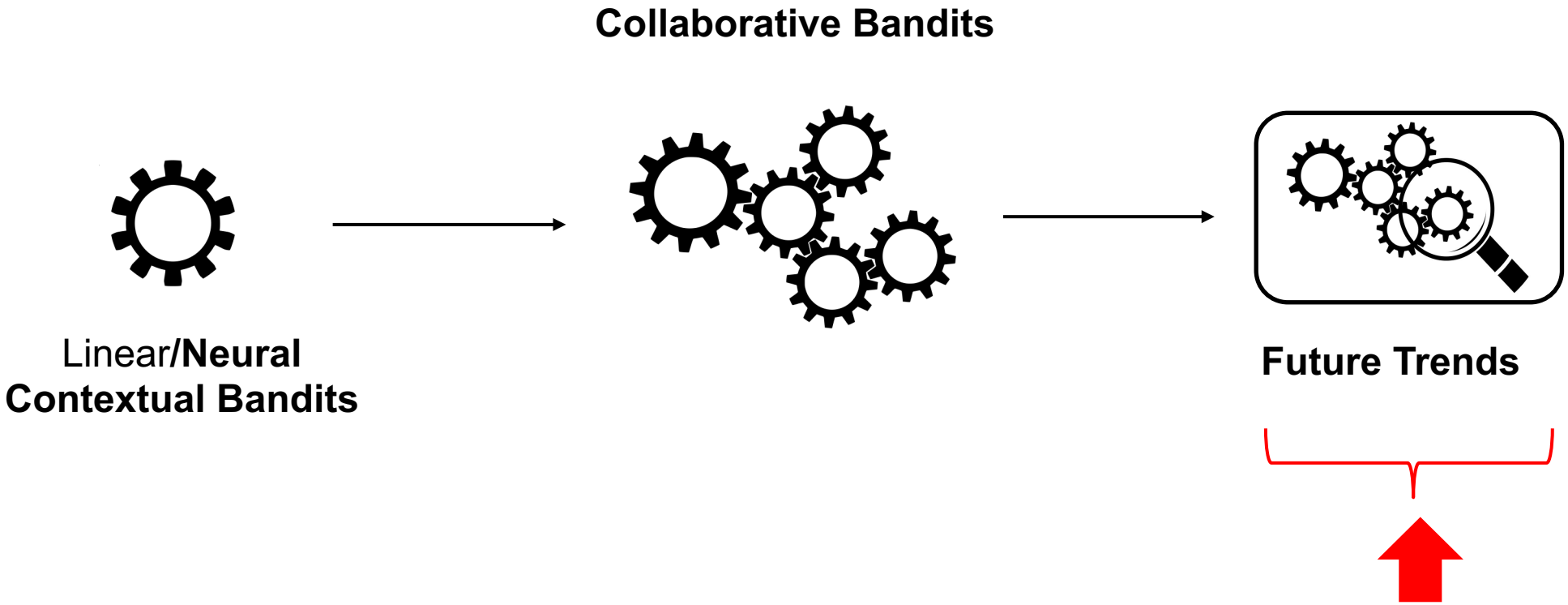
Figure: Regret comparison on Mnist+NotMnist with different reward availability.

Observation:

- **With one sub-reward, MuFasa still outperforms all baselines.**
- **Without any sub-reward, MuFasa's performance is close to the best baseline.**

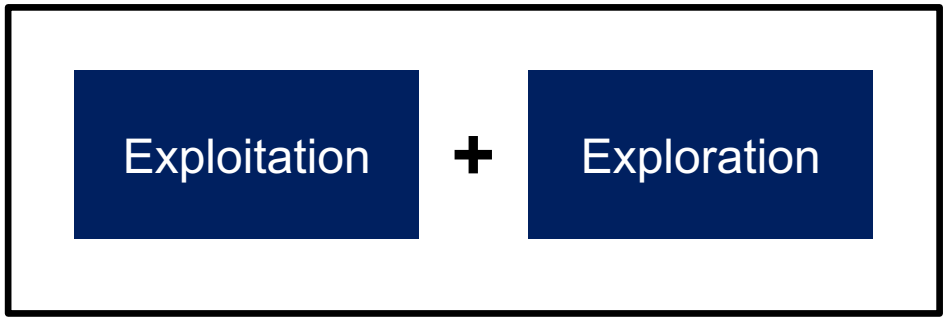
- Yikun Ban et. al., Multi-facet contextual bandits: A neural network perspective. KDD 2022.

Roadmap


Collaborative Bandits
Linear/Neural
Contextual Bandits
Future Trends

Trustworthy Exploration: Transparency

Q: Can we have a **transparent** exploration with clear rationales and explanations?



➢ **Challenges:**

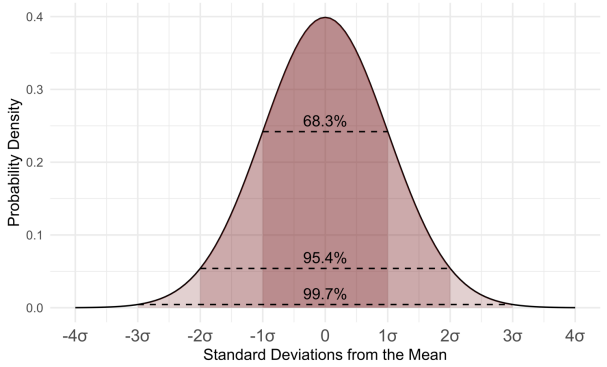
- ❑ More exploration models based on neural networks (**Black Box**).

Black Box !

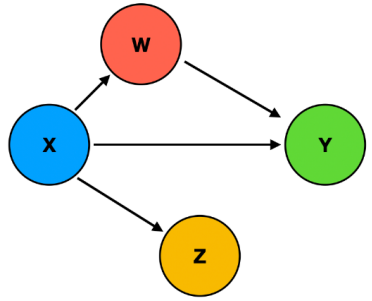
E.g. [Ban et al. ICLR 2022]

➢ **Future Directions:**

- ❑ Bayesian Bandits/RL.
- ❑ Causal Bandits/RL.



Statistics

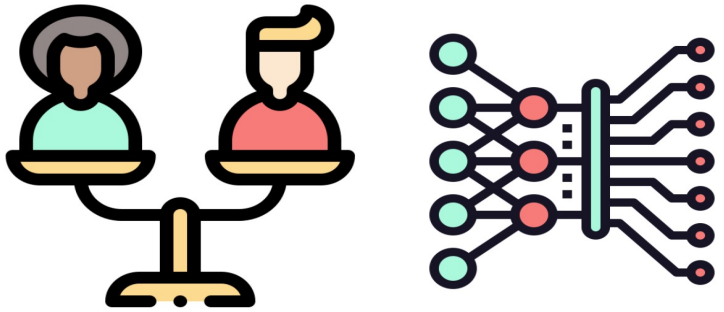


Causal Inference

• Castelvecchi, Davide. "Can we open the black box of AI?." *Nature News* 538.7623 (2016): 20.

Trustworthy Exploration: Fairness

Q: How to ensure **fairness** in the context of exploration?



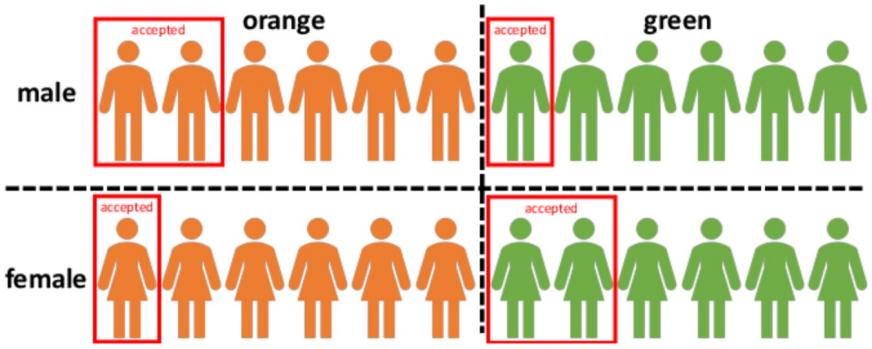
Group Fairness

➢ **Challenges:**

- ❑ **Non-IID** data.
- ❑ balance required between **exploration power** and **fairness**.

➢ **Future Directions:**

- ❑ Derive fairness confidence interval for exploration.
- ❑ Fairness Regularization.



Group Fairness [1]

- Joseph, Matthew, et al. "Fairness in learning: Classic and contextual bandits." NeurIPS 2016.
- Kang, Jian, et al. "Multifair: Multi-group fairness in machine learning." Bigdata 2021.

Trustworthy Exploration: Privacy

Q: Can we have an exploration strategy preserving privacy?

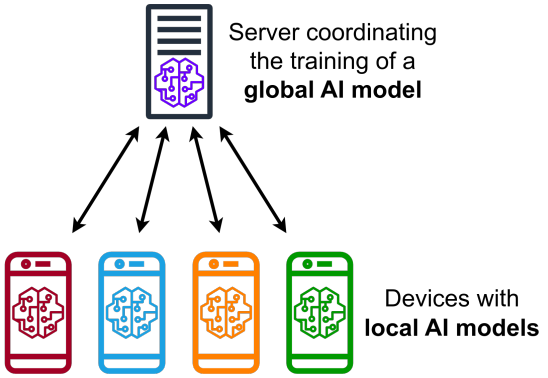
User Privacy

➢ **Challenges:**

- ❑ Privacy-preserving exploration methods.

➢ **Future Directions:**

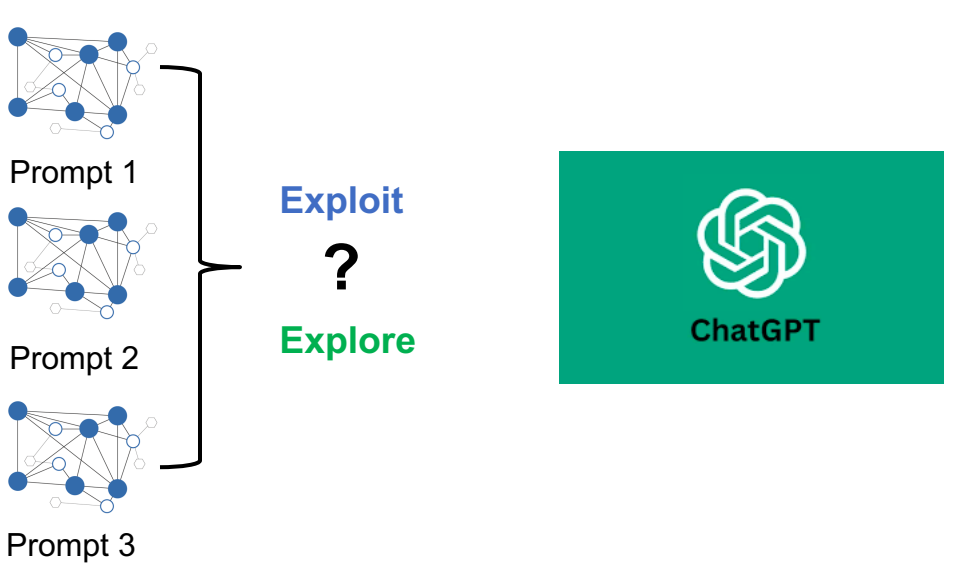
- ❑ Federated Bandits/RL.



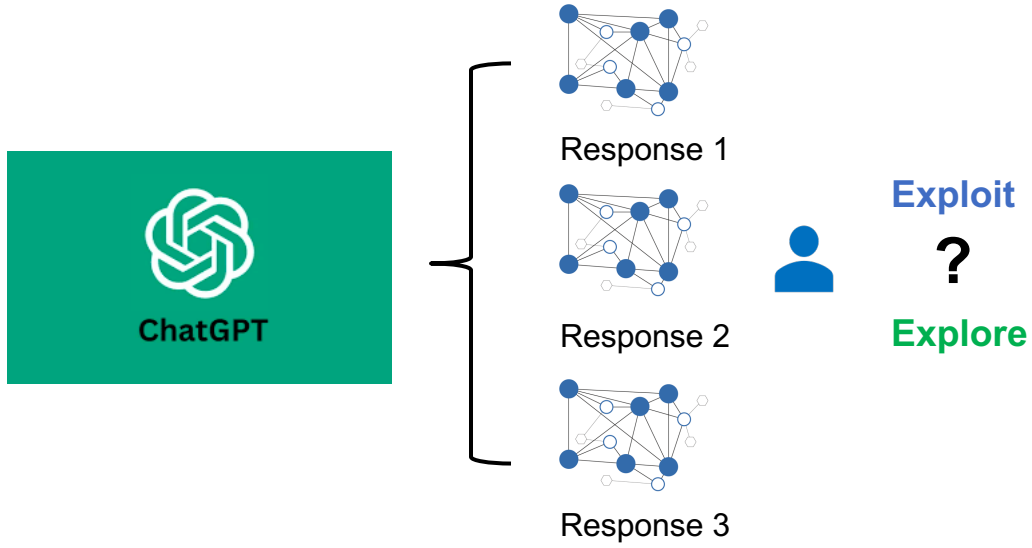
Federated Learning

- Zhang, Chen, et al. "A survey on federated learning." *Knowledge-Based Systems* 216 (2021): 106775.
- Dai, Zhongxiang, et al. "Federated neural bandits." ICLR 2023.

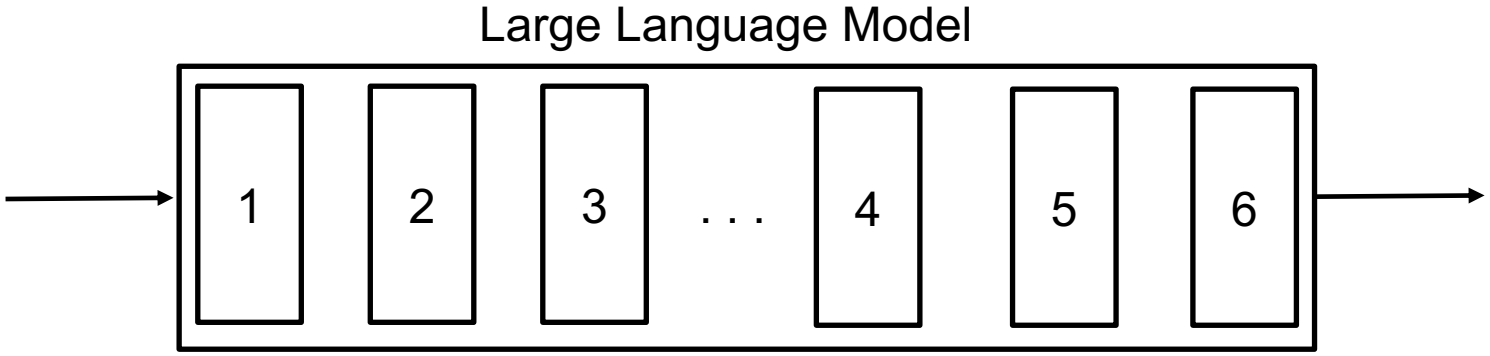
Customized Exploration: Large Language Model



(1) Prompting with Exploration



(2) Answering with User-specific Exploration

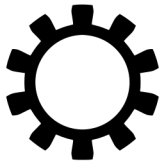


(3) Fine-tuning with Exploration

- Lin, Xiaoqiang, et al. "Use your INSTINCT: instruction optimization using neural bandits coupled with transformers." ICML 2024.
- Chen, Zekai, et al. "Online Personalizing White-box LLMs Generation with Neural Bandits." *arXiv preprint arXiv:2404.16115* (2024).
- Köpf, Andreas, et al. "Openassistant conversations-democratizing large language model alignment." NeurIPS 2023.
- Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." NeurIPS 2023.
- Zhang, Qingru, et al. "Platon: Pruning large transformer models with upper confidence bound of weight importance." ICML 2022.

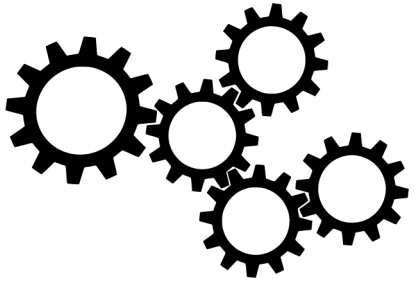
Roadmap

- NeuralUCB
- NeuralTS
- EE-Net
- Neural-LinUCB
- Neural-Bandit-Perturbed Reward
- Neural-SquareCB



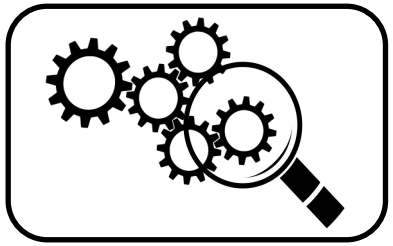
Linear/**Neural**
Contextual Bandits

Collaborative Bandits



- Clustering of Linear Bandits
- Clustering of Neural Bandits
- Graph Bandit Learning
  - Graph Neural Bandits
  - Arm Group Graph
- Multi-facet Neural Bandits

- Trustworthy Neural Bandits with Recommendation
- LLM with Bandits for Exploration



Future Trends

THE WEB CONFERENCE ACM

The Grainger College of Engineering
IBM-Illinois Discovery Accelerator Institute

AIFARMS
Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

NSF

Neural Contextual Bandits for Personalized Recommendation

Yikun Ban **Yunzhe Qi** **Jingrui He**

University of Illinois Urbana-Champaign

{Yikunb2, Yunzheq2, Jingrui}@illinois.edu

Time: 9:00 AM – 12:30 PM, 13 May 2024

Location: Virgo 1, Resorts World Sentosa Convention Centre, Singapore

Website: www.banyikun.com/wwwtutorial/

ILLINOIS