Towards Fairness in Transportation Gig Markets: Identifying, Imitating, and Mitigating Algorithm Discrimination via Deep Reinforcement Learning

Zijian Zhao (zzhaock@connect.ust.hk)

The Hong Kong University of Science and Technology

Association for the Advancement of Artificial Intelligence

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Motivation & Background

On-demand food delivery platforms have begun utilizing Reinforcement Learning (RL) techniques for labor management. While these methods lead to enhanced system performance and efficiency, they also raise concerns about **algorithmic discrimination** stemming from the **overuse of individual data**. For instance, platforms may implement personalized payment schemes for couriers based on varying priorities during order assignment. **Data Privacy Regulations (DPR)** such as GDPR, CCPA, and PIPL aim to address these concerns by granting couriers the right to decide whether to share their personal data. However, most current studies focus primarily on statistical analyses and user studies, leaving key questions unanswered:

- **How do discriminatory algorithms impact both platforms and couriers?**
- **What effect do DPR policies have on the food delivery market?**

Problem Setup

Platform (Hybrid-Action MAMDP):

- Assign orders to couriers (discrete action) and bundle them with en-route orders.
- Determine payments for each courier-order pair (continuous action).
- Differentiate couriers based on their historical behavior data.

Couriers (Two Time-Scale Decision):

- **Long-term Decision (CMAB):**
  - Opt-in: Work and share individual data.
  - Opt-out: Work but not share individual data.
  - Exit: Do not work.
- **Short-term Decision (Logit Model):** Assess whether to accept the assigned order with the proposed payment from the platform.
- **Reservation Value (Core Concept):** The minimum compensation a courier is willing to accept for their time and effort.

Customers (Environment):

- Place orders at arbitrary times.
- Decline orders if not confirmed within a specified maximum waiting time.

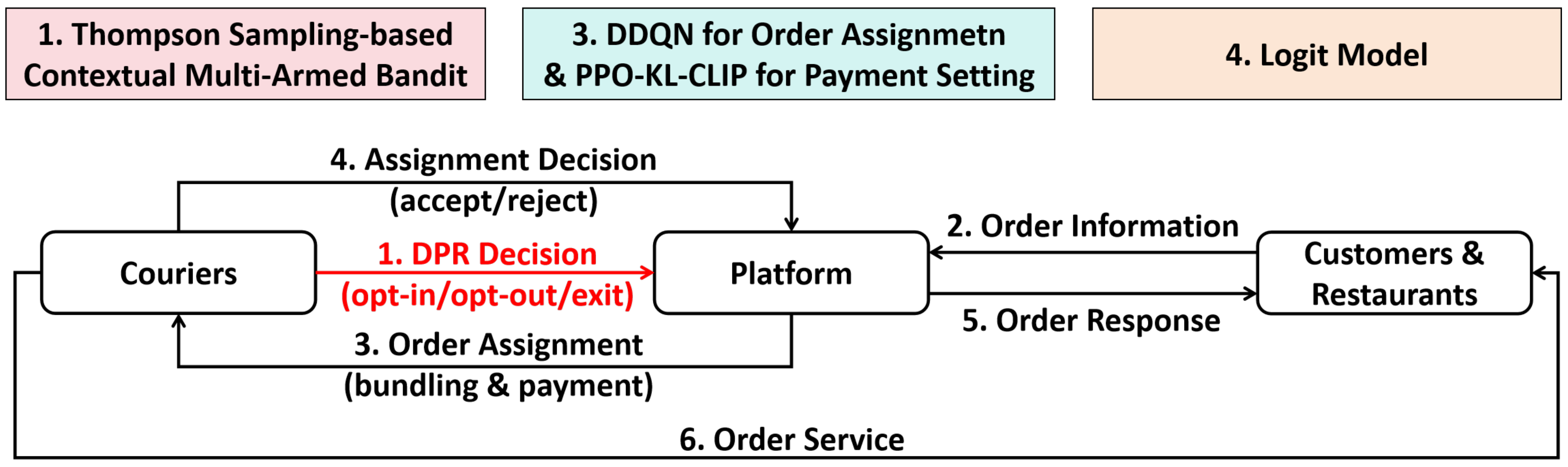
Conclusion

- We propose a DDQN-PPO method for hybrid action control to model platform behavior.
- We introduce a MA-CMAB using a neural Thompson sampling method to model courier behavior.
- **Insights:** Data privacy regulation fosters a win-win scenario for all stakeholders in the food delivery market. By offering more working opportunities for opt-out couriers, a higher number of them choose to work on the platform. This increase in active couriers enables the platform to generate higher income by fulfilling more orders, which in turn reduces delivery times and the overtime rate for customers [1, 2].

References

[1] Zijian Zhao et al. "Discriminatory Order Assignment and Payment-Setting of On-Demand Food-Delivery Platforms: A Multi-Action and Multi-Agent Reinforcement Learning Framework". In: *Transportation Research Part E: Logistics and Transportation Review* 208 (2026), p. 104653.

[2] Zijian Zhao et al. "The impacts of data privacy regulations on food-delivery platforms". In: *Transportation Research Part C: Emerging Technologies* 181 (2025), p. 105364.



1. Thompson Sampling-based Contextual Multi-Armed Bandit

3. DDQN for Order Assignmetn & PPO-KL-CLIP for Payment Setting

4. Logit Model

Couriers

Platform

Customers & Restaurants

4. Assignment Decision (accept/reject)

1. DPR Decision (opt-in/opt-out/exit)

3. Order Assignment (bundling & payment)

2. Order Information

5. Order Response

6. Order Service

Methodology

Platform MAMDP:

- Employ IDDQN for order assignment and IPPO for payment setting.
- Integrate IDDQN and IPPO by using the Q-Network from IDDQN as the critic for IPPO.

$$V_{\pi^P}^{PPO}(s_{i,t}|\pi^A) = \mathbb{E}_{\pi^A,\pi^P}\left[\sum_{\tau\in\mathcal{T}}\gamma^\tau\cdot\mathcal{R}_i(s_{i,t+\tau},a_{i,t+\tau})\mid s_{i,t}\right]$$
$$= \mathbb{E}_{\pi^A,\pi^P}\left[\sum_{\tau\in\mathcal{T}}\gamma^\tau\cdot\mathcal{R}_i(s_{i,t+\tau},a_{i,t+\tau})\mid s_{i,t},\pi^A(s_{i,t})\right]$$
$$= Q_{\pi^A}^{DDQN}(s_{i,t},\kappa_{i,t}^*|\pi^P)$$

- The joint policy $\pi = [\pi^A, \pi^P]$ consists of the assignment policy and payment policy.
- The joint action $a = [\kappa, p]$ comprises assignment action and payment action.

Courier MA-CMAB:

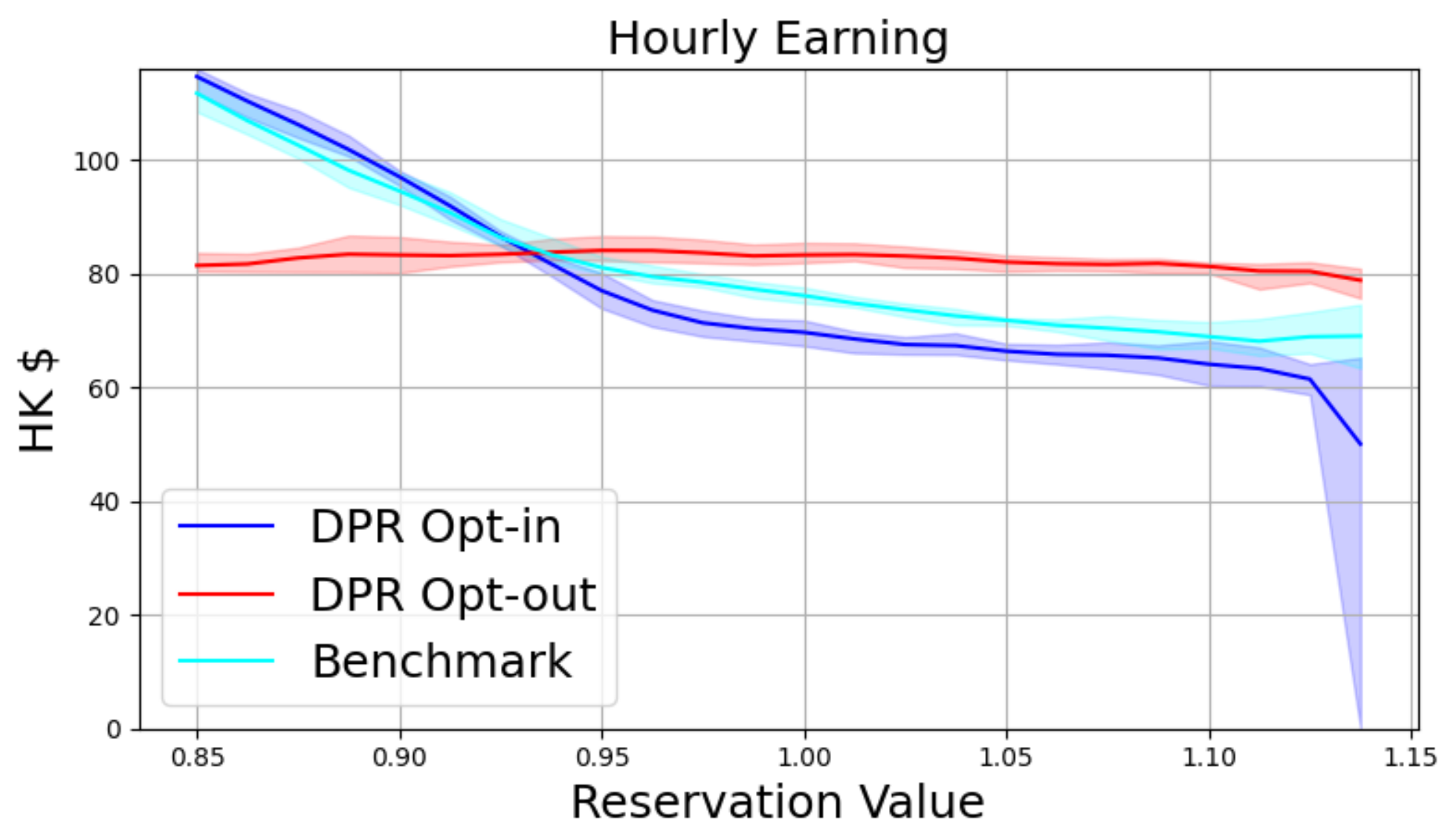
- Model the reward of each bandit as a Gaussian distribution.
- Generate the distribution using the neural Thompson sampling method.
- Optimize the network via maximum likelihood estimation (MLE).

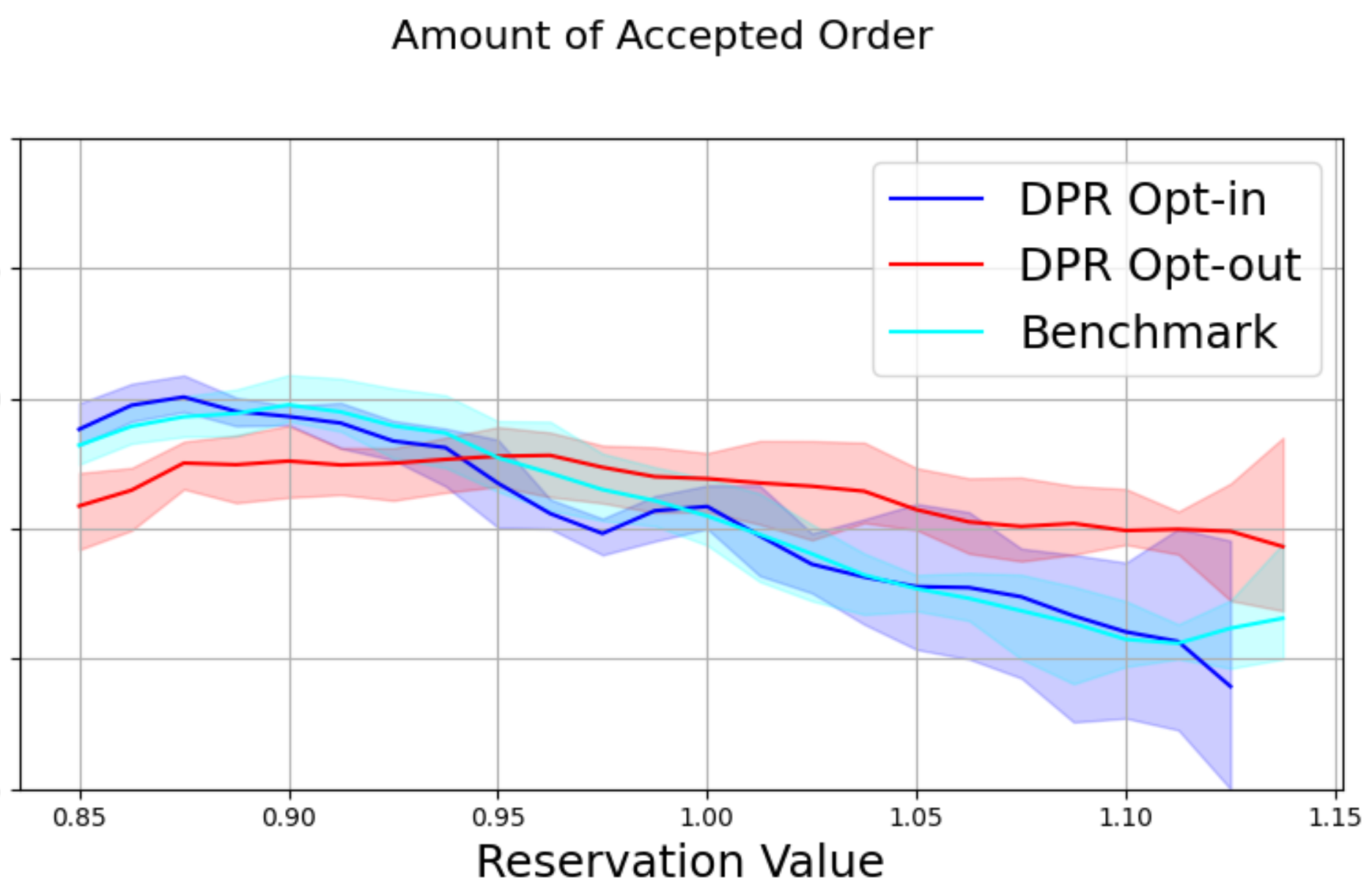
$$\Phi = \arg\max_\Phi P(y|N(\mu(x;\Phi),\sigma(x;\Phi)^2)) \ ,$$
$$L_{mle}(\{\mu,\sigma\},y) = -\log P(y|N(\mu,\sigma^2))$$
$$= \log(\sigma) + \frac{(y-\mu)^2}{\sigma}$$

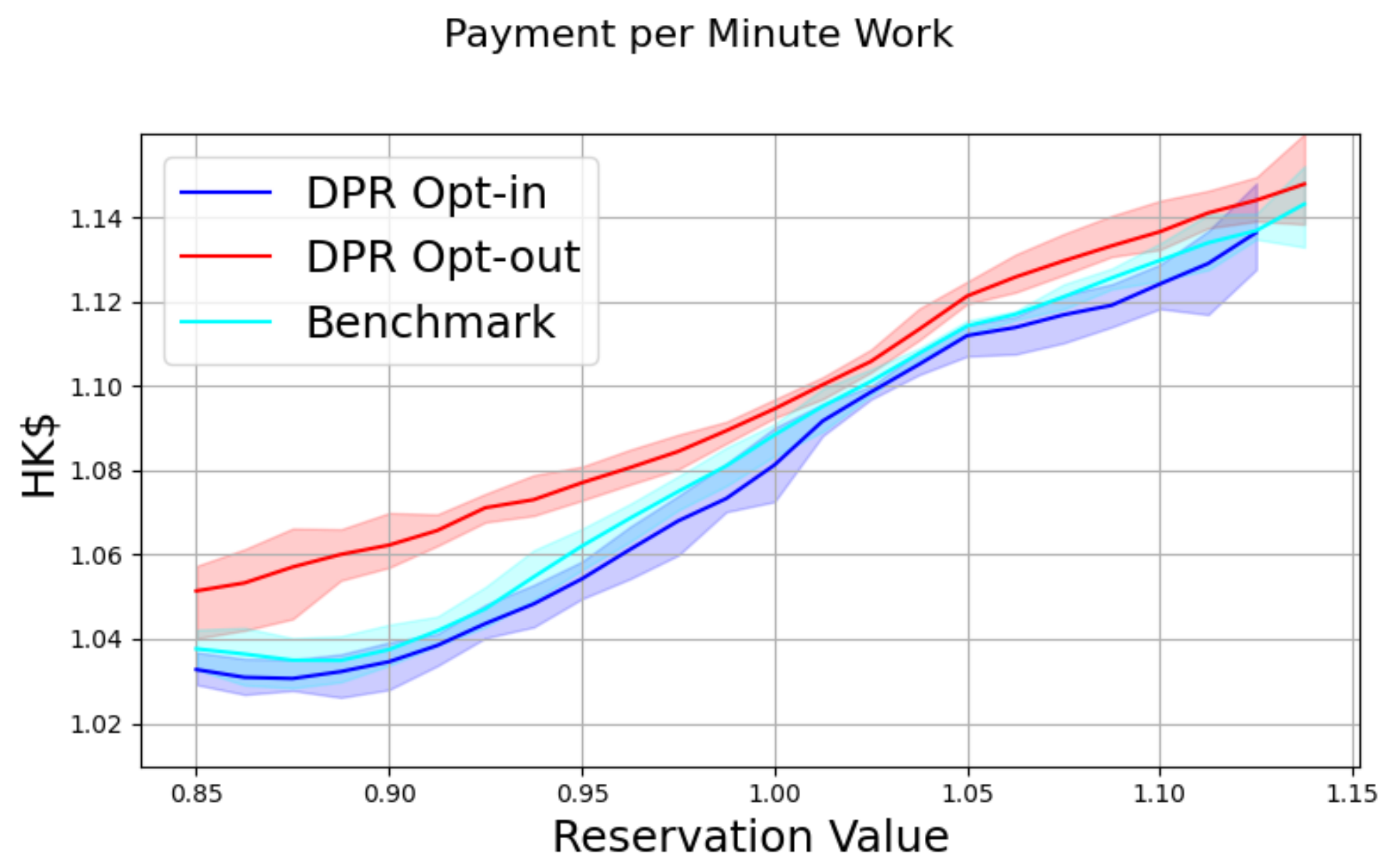
- The network $N(\mu(x;\Phi),\sigma(x;\Phi)^2)$ outputs a Gaussian distribution for each bandit.
- Variables x, y, and Φ represent input, ground truth, and network parameters, respectively.

Case Study

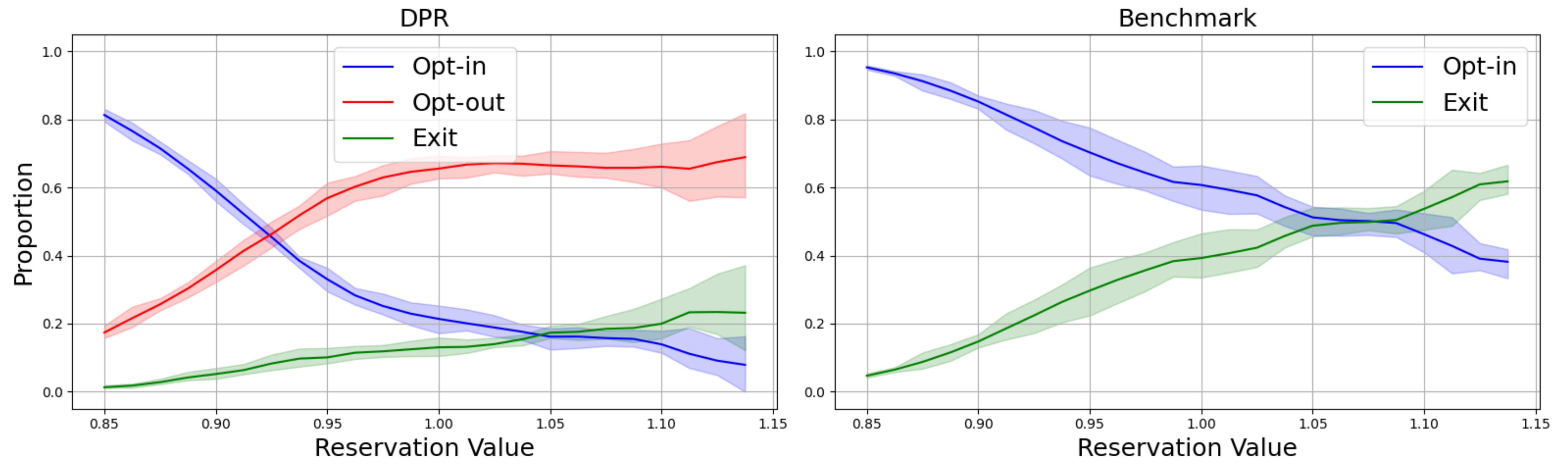
- **DPR:** Couriers can choose to opt-in, opt-out, or exit.
- **Benchmark:** Couriers can choose to opt-in or exit.
- **Results:** The platform favors low-reservation opt-in couriers by offering more work opportunities; however, it struggles to distinguish between opt-out couriers.







Policy	Platform_Profits	Active_Courier_Rate	Order_Service_Rate	Delivery_Time	Order_Overtime_Rate
DPR	12,310 HK$	90.64%	57.18%	20.70 min	5.72%
Benchmark	11,170 HK$	72.84%	45.57%	21.17 min	6.45%



Courier Choice