Appendix: Personalization, Consumer Search and Algorithmic Pricing

Liying Qiu Yan Huang Param Vir Singh Kannan Srinivasan*

A.1 Optimal Search

When $n = 2$, the search process can be simplified as follows. Since consumers incur no cost to inspect the product ranked in the first position, they will only need to make a decision on whether to search the product ranked in the second position, i.e., $d_{i(1)}$. When consumer i is making this decision, the current optimal value is $u^*_{i(2)} = \max\{u_{i(0)}, u_{i(1)}\}$.

Given $u^*_{i(2)}$, we can calculate the expected marginal gain for consumer i from searching the product at position 2 as follows:

$$B(u^*_{i(2)}) = \int_{u^*_{i(2)}}^{\infty} (u_{i(2)} - u^*_{i(2)}) f(u_{i(2)}|\psi_i) du_{i(2)}, \tag{A1}$$

where $f(u_{i(2)}|\psi_i)$ is the posterior probability density function of $u_{i(2)}$ after observing the ranking ψ_i) and realized utility for the top-ranked product $u_{i(1)}$.

Let $z_{i(2)}$ be the value ("reservation value") such that the consumer is indifferent between obtaining utility $z_{i(2)}$ immediately (which saves additional search costs, ρ) or evaluating seller 2's product which gives her an option to choose between $z_{i(2)}$ and $u_{i(2)}$:

$$\rho = B(z_{i(2)}) = \int_{z_{i(2)}} (u_{i(2)} - z_{i(2)}) f(u_{i(2)}|\psi_i) du_{i(2)} \tag{A2}$$

When the best value in the consideration set is lower than the reservation value, consumer i will continue to search for product at position 2; otherwise she will stop searching.

*All authors are at Carnegie Mellon University.

The probability that consumer i searches only the product ranked in the position 1 but not the product in position 2 is $\mathbb{P}[u^*_{i(2)} \geq z_{i(2)}]$; the probability that consumer i searches both products is $\mathbb{P}[u^*_{i(2)} < z_{i(2)}]$. The conditional probability that consumer i will purchase the product at position h given that the consumer evaluates only the product in position 1 is

$$\eta_{i(h)} = \mathbb{P}[u_{i(h)} = \max\{u_{i(0)}, u_{i(1)}\} | d_{i(1)} = 0] \text{ where } h \in \{0, 1\} \tag{A3}$$

The same conditional probability given that the consumer evaluates both product is

$$\eta_{i(h)} = \mathbb{P}[u_{i(h)} = \max\{u_{i(0)}, u_{i(1)}, u_{i(2)}\} | d_{i(1)} = 1] \text{ where } h \in \{0, 1, 2\} \tag{A4}$$

Note that $d_{i(1)}$ simplifies to a binary decision in this two-product case, as consumers only have one remaining product to search after inspecting the first, eliminating the need to decide which product to search.

A.2 Demand and Consumer Surplus

A.2.1 Unpersonalized Ranking

The demand for the product at position 1 under unpersonalized ranking is given as:

$$D_{(1)} = \int_{z_{i(2)}}^{\infty} F_{u_{i(0)}}(t)(1 - F_{u_{i(2)}}(t)) f_{u_{i(1)}}(t) dt + \int_{\max\{u_{i(0)}, u_{i(2)}\}}^{\infty} f_{u_{i(1)}}(t) dt \tag{A5}$$

where $u_{i(r)}$ denotes the utility at position r and $u_{i(0)}$ represents the utility from outside option for consumer i. $F_{u_{i(k)}}$ denotes the CDF of the random variable $u_{i(k)}$, and $f_{u_{i(k)}}$ denotes the PDF of the random variable $u_{i(k)}$.[1] The first part on the right-hand side represents the demand of first-ranked product when the utility of the product in the second position is higher than the first one for consumer i but consumer i decides not search the product in the second position and purchase the first one. Mathematically, it requires $u_{i(0)} < u_{i(1)} < u_{i(2)}$ and $u^*_{i(2)} > z_{i(2)}$. Since $u^*_{i(2)} = \max\{u_{i(0)}, u_{i(1)}\}$ and $u_{i(0)} < u_{i(1)}$, $u^*_{i(2)} > z_{i(2)}$ is equivalent to $u_{i(1)} > z_{i(2)}$. The second part on the right-hand side

[1] The notations are slightly adjusted from those used in the body of the paper for simplicity.

represents the demand when the utility of the product on the first position is the highest hence consumer i will purchase the first one regardless of their searching decision.

The demand for the product ranking in the second position only comes from the case when consumer i decides to inspect the product in the second position and the second ranked product yields the higher utility for consumer i than the first-ranked product and outside option. It can be expressed as

$$D_{(2)} = \int_{-\infty}^{z_{i(2)}} (1 - F_{u_{i(2)}}(t)) f_{u^*_{i(2)}}(t) dt \tag{A6}$$

The condition above requires that $u^*_{i(2)} < z_{i(2)}$ so that consumer i will search for the second ranked product, and at the same time, the value of $u_{i(2)}$ is greater than both $u_{i(0)}$ and $u_{i(1)}$, which is equivalent to $u_{i(2)} > u^*_{i(2)}$.

The consumer surplus under unpersonalized ranking (indicated by the superscript U) is given as

$$\begin{aligned}
\mathbb{E}^U(CS) = & \int_{z_{i(2)}}^{\infty} t f_{u^*_{i(2)}}(t) dt \\
& + \int_{-\infty}^{z_{i(2)}} f_{u^*_{i(2)}}(t) [\int_{t}^{\infty} s f_{u_{i(2)}}(s) ds - \rho] dt \\
& + \int_{-\infty}^{z_{i(2)}} f_{u^*_{i(2)}}(t) [\int_{-\infty}^{t} t f_{u_{i(2)}}(s) ds - \rho] dt
\end{aligned} \tag{A7}$$

The first part of the right-hand side of the equation corresponds to the case where consumer i does not inspect the product in the second position and chooses the option that yields a higher utility between the outside option and the product in the first position without incurring any further search cost. The second part corresponds to the case where consumer i inspects the product in the second position and finds that it generates the highest utility. The third part corresponds to the case where consumer i inspects the second product and finds that it does not provide any higher utility than the first product or the outside option. As a result, the consumer chooses the better of the first product or the outside option.

A.2.2 Personalized Ranking

Under perfect personalized ranking where the accuracy of ranking is 100%, consumers realize that the platform perfectly knows their match values and the utilities associated with each product, making it optimal for them to search only the top-ranked product. This is because among all products, they derive the highest utility from the top-ranked product. The demand for product at the top position under perfect personalized ranking is $1 - F_{u_{i(1)}}(\max\{u_{i(0)}, u_{i(2)}\})$; there is no demand for product at the second position. Therefore, the demand for product j can be expressed as

$$D_j^P = \frac{1}{1 + \exp(\frac{\mu \ln(\exp(\frac{a_0}{\mu}) + \exp(\frac{a_{-j} - \phi p_{-j}}{\mu})) - a_j + \phi p_j}{\mu})} \tag{A8}$$

where $-j$ indicates the competitor of product j.

Since consumers incur no search cost under perfect personalized ranking, the consumer surplus is the expected value of the maximum utility among all products, which can be defined as

$$\mathbb{E}^P(CS) = \gamma + \log(1 + \sum_j e^{a_j - \phi p_j}) \tag{A9}$$

Under imperfect personalized ranking where the accuracy of the ranking is below 100%, after observing the product rank and the realized match values for the products that have been searched, consumers will update their belief about the match values for the un-searched products, and choose the next product to search based on the updated belief. To see this, consider a duopoly case. Let us use $\boldsymbol{J} =< 0, j, k >$ to denote a ranking where product j ranks first and product k ranks second, and $\boldsymbol{K} =< 0, k, j >$ to denote a ranking where product k ranks first and product j ranks second. The accuracy of the ranking, λ, defined as

$$\begin{aligned} \lambda &= P(\boldsymbol{J}|u_{ij} > u_{ik}) = P(\boldsymbol{K}|u_{ij} < u_{ik}) \\ 1 - \lambda &= P(\boldsymbol{K}|u_{ij} > u_{ik}) = P(\boldsymbol{J}|u_{ij} < u_{ik}) \end{aligned} \tag{A10}$$

Let ψ^{IP} denote the event that the observed ranking. The posterior belief about the product that has not yet been searched (second-ranked) becomes

$$f(\epsilon_{i(2)}|\psi^{IP}, \epsilon_{i(1)}) = \frac{\lambda f(\epsilon_{i(2)}|\epsilon_{i(2)} < \Delta_{i(1)})F_{\epsilon_{i(2)}}(\Delta_{i(1)}) + (1-\lambda)f(\epsilon_{i(2)}|\epsilon_{i(2)} > \Delta_{i(1)})(1 - F_{\epsilon_{i(2)}}(\Delta_{i(1)}))}{\lambda F_{\epsilon_{i(2)}}(\Delta_{i(1)}) + (1-\lambda)(1 - F_{\epsilon_{i(2)}}(\Delta_{i(1)}))} \tag{A11}$$

where $\Delta_{i(1)} = a_{(1)} - \phi p_{(1)} + \epsilon_{i(1)} - a_{(2)} + \phi p_{(2)}$, $f(\epsilon_{i(2)}|\cdot)$ denotes the conditional probability density function of $\epsilon_{i(2)}$, and $F_{\epsilon_{i(2)}}$ denotes the cumulative distribution function of $\epsilon_{i(2)}$. Thus, under imperfect personalized ranking, we use this posterior belief about the non-top-ranked product to re-calculate the search gain. This posterior search gain can be expressed as

$$B(u^*_{i(2)}) = \int_{\Delta^*_{i(1)}}^{\infty} \frac{(1-\lambda)f(\epsilon_{i(2)}|\epsilon_{i(2)} > \Delta_{i(1)})(1 - F_{\epsilon_{i(2)}}(\Delta_{i(1)}))}{\lambda F_{\epsilon_{i(2)}}(\Delta_{i(1)}) + (1-\lambda)(1 - F_{\epsilon_{i(2)}}(\Delta_{i(1)}))}(\epsilon_{i(2)} - \Delta^*_{i(1)})d\epsilon_{i(2)} \tag{A12}$$

When there are two firms, consumers decide whether to continue searching the second-ranked product after inspecting the top-ranked one, based on this posterior search gain.

To compute the expected demand and consumer surplus under imperfect personalized ranking, we first separately consider the expected demand and consumer surplus when the ranking is correct and when it is incorrect. We start with the case when the ranking is correct. The demand for product j under correct ranking (denoted as superscript C) is the same as D^P_j. That is

$$D^C_j = D^P_j$$

. Define the event $\mathbb{A}(\boldsymbol{J}) = \{u_{i0} > u_{ij} > u_{ik}|\boldsymbol{J}\}$ and $\mathbb{A}^C(\boldsymbol{J}) = \{u_{ij} > \max\{u_{i0}, u_{ik}\}|\boldsymbol{J}\}$. Similarly, define the event $\mathbb{A}(\boldsymbol{K}) = \{u_{i0} > u_{ik} > u_{ij}|\boldsymbol{K}\}$ and $\mathbb{A}^C(\boldsymbol{K}) = \{u_{ik} >$

$\max\{u_{i0}, u_{ij}\}|\boldsymbol{K}\}$. The expected consumer surplus under ranking $\boldsymbol{J}$ can be written as

$$\begin{aligned}\mathbb{E}^C(CS) = &\int u_{i0}\mathbb{1}(\boldsymbol{J})\mathbb{1}(\mathbb{A}(\boldsymbol{J}))f_{\boldsymbol{u}_i}(t)dt - \rho\int \mathbb{1}(\boldsymbol{J})\mathbb{1}(\mathbb{A}(\boldsymbol{J}))\mathbb{1}(u_{i0} < z^{\boldsymbol{J}}_{i(2)})f_{\boldsymbol{u}_i}(t)dt \\ &+ \int u_{i(1)}\mathbb{1}(\boldsymbol{J})\mathbb{1}(\mathbb{A}^C(\boldsymbol{J}))f_{\boldsymbol{u}_i}(t)dt - \rho\int \mathbb{1}(\boldsymbol{J})\mathbb{1}(\mathbb{A}^C(\boldsymbol{J}))\mathbb{1}(u_{i(1)} < z^{\boldsymbol{J}}_{i(2)})f_{\boldsymbol{u}_i}(t)dt \\ &+ \int u_{i0}\mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}(\boldsymbol{K}))f_{\boldsymbol{u}_i}(t)dt - \rho\int \mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}(\boldsymbol{K}))\mathbb{1}(u_{i0} < z^{\boldsymbol{K}}_{i(2)})f_{\boldsymbol{u}_i}(t)dt \\ &+ \int u_{i(1)}\mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}^C(\boldsymbol{K}))f_{\boldsymbol{u}_i}(t)dt - \rho\int \mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}^C(\boldsymbol{K}))\mathbb{1}(u_{i(1)} < z^{\boldsymbol{K}}_{i(2)})f_{\boldsymbol{u}_i}(t)dt\end{aligned} \tag{A13}$$

Here, $z^{\boldsymbol{J}}_{i(2)}$ and $z^{\boldsymbol{K}}_{i(2)}$ are the reservation value defined previously using Equations A2 and A12 under ranking $\boldsymbol{J}$ and $\boldsymbol{K}$ respectively. The first item of the first line on the right-hand side denotes the expected utility under ranking $\boldsymbol{J}$ and utility combinations $\mathbb{A}(\boldsymbol{J})$, and the second item denotes the search cost under ranking $\boldsymbol{J}$ and utility combinations $\mathbb{A}(\boldsymbol{J})$. Note that only when $u_{i0} < z^{\boldsymbol{J}}_{i(2)}$, consumers will inspect the second ranked product in this case. Similarly, the first and second items in the second (third, fourth) line on the right-hand side denote respectively the expected utility and search cost under ranking $\boldsymbol{J}$ and utility combinations $\mathbb{A}^C(\boldsymbol{J})$ (ranking $\boldsymbol{K}$ and utility combinations $\mathbb{A}(\boldsymbol{K})$, ranking $\boldsymbol{K}$ and utility combinations $\mathbb{A}^C(\boldsymbol{K})$).

Let us now move on to the case where the ranking is incorrect. The demand for product j under incorrect ranking (denoted as superscript I) can be given as

$$D^I_j = \int_{-\infty}^{z_{i(2)}} (1 - F_{u_{ij}}(t))f_{\max\{u_{i0},u_{ik}\}}(t)dt + \int_{z_{i(2)}}^{\infty} F_{u_{i0}}(t)(1 - F_{u_{ik}}(t))f_{u_{ij}}(t)dt \tag{A14}$$

where u_{ik} denotes the utility for individual i from the competitor of product j, which is different from $u_{i(k)}$. The first part on the right-hand side indicates the demand when product j's utility is higher than product k but product j is incorrectly ranked in the second position. In this case, consumer i will purchase product j when searching for the second position (i.e., $\max\{u_{i0}, u_{ik}\} < z_{i(2)}$) and finding j gives the highest utility (i.e., $u_{ij} > \max\{u_{i0}, u_{ik}\}$). In contrast, the second part of the right side of the equation indicates the demand when the product j's utility is lower than product k (i.e., $u_{ik} > u_{ij}$)

but product j is incorrectly ranked on the top position. In this case, consumer i will choose product j if she decides not to search the second position (i.e., $\max\{u_{i0}, u_{ij}\} > z_{i(2)}$) and finds product j provides a higher utility than the outside option (i.e., $u_{i0} < u_{ij}$).

If the ranking is incorrect, the expected consumer surplus under personalized ranking for ranking $\boldsymbol{K}$ is

$$\begin{aligned}\mathbb{E}^I_{\boldsymbol{K}}(CS) = &\int u_{i0}\mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}(\boldsymbol{J}))f_{\boldsymbol{u}_i}(t)dt - \rho\int \mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}(\boldsymbol{J}))\mathbb{1}(u_{i0} < z^{\boldsymbol{K}}_{i(2)})f_{\boldsymbol{u}_i}(t)dt \\ &+ \int (u_{ij}-\rho)\mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}^C(\boldsymbol{J}))\mathbb{1}(\max\{u_{i0}, u_{ik}\} < z^{\boldsymbol{K}}_{i(2)})f_{\boldsymbol{u}_i}(t)dt \\ &+ \int \max\{u_{i0}, u_{ik}\}\mathbb{1}(\boldsymbol{K})\mathbb{1}(\mathbb{A}^C(\boldsymbol{J}))\mathbb{1}(\max\{u_{i0}, u_{ik}\} > z^{\boldsymbol{K}}_{i(2)})f_{\boldsymbol{u}_i}(t)dt \qquad \text{(A15)}\end{aligned}$$

The first line on the right-hand side shows the expected consumer surplus when $u_{i0} > u_{ij} > u_{ik}$ (i.e., the true state is $\mathbb{A}(\boldsymbol{J})$) but due to error in ranking, the observed ranking is $\boldsymbol{K} =< 0, k, j >$. In this case, the maximum utility is u_{i0} regardless of searching, and the search cost is incurred only in the presence of searching (i.e., $u_{i0} < z^{\boldsymbol{K}}_{i(2)}$). The second line on the right-hand side shows the expected consumer surplus when $u_{ij} > \max\{u_{i0}, u_{ik}\}$ (i.e., the true state is $\mathbb{A}^C(\boldsymbol{J})$). In this case, the maximum utility will be u_{ij} if consumer i searches the second product (i.e., $\max\{u_{i0}, u_{ik}\} < z^{\boldsymbol{K}}_{i(2)}$)); otherwise, as shown in the third line, consumer i's utility will be $\max\{u_{i0,}u_{ik}\}$ without further searching. $\mathbb{E}^I_{\boldsymbol{J}}(CS)$ can be similarly defined. Therefore the expected consumer welfare from incorrect personalized ranking is

$$\mathbb{E}^I(CS) = \mathbb{E}^I_{\boldsymbol{J}}(CS) + \mathbb{E}^I_{\boldsymbol{K}}(CS) \qquad \text{(A16)}$$

So far, we have considered two extreme cases — (1) rankings are always correct and (2) rankings are always incorrect. Consider a scenario of imperfect ranking, where the prediction accuracy of the personalized ranking is λ. The expected demand of product j is the weighted average of correct ranking and incorrect ranking, where the weight is the

ranking algorithm accuracy λ:

$$D_j^{IP} = (1-\lambda)D_j^I + \lambda D_j^C \tag{A17}$$

Similarly, the expected consumer welfare under imperfect personalized ranking is

$$\mathbb{E}^{IP}(CS) = (1-\lambda)\mathbb{E}^I(CS) + \lambda \mathbb{E}^C(CS) \tag{A18}$$

A.3 Platform's Information and Ranking Systems

We assume that platforms possess significantly accurate information about consumers, as they have access to vast amount of consumer data, such as browsing behavior, purchase history, and interactions. With this data, platforms can more accurately infer consumer preferences, price elasticity, and seller (firm) quality (i.e., vertical differentiation). In our paper, we assume a stable market condition where those parameter values do not change over time, and the platform has collected enough data to estimate these parameters. Since our focus is to examine the prices that competing online learning pricing algorithms converge to, we treat the platform's estimates and the ranking rules based on them as stable and exogenously given, without modeling how the platform updates its estimates through online learning. This assumption is realistic, as platforms like Amazon typically have far more information about consumers than individual sellers (excluding Amazon itself).

Furthermore, the platform's ranking algorithm—whether personalized or unpersonalized—is exogenously determined. The value of λ, which captures accuracy of personalized ranking, is assumed to be common knowledge and fixed. Importantly, we do not model the platform's actions as strategic, nor do we solve for an optimal choice of λ. This assumption allows us to focus on the competition between firms under different ranking regimes rather than on the platform's decision-making process. The platform does not engage in profit-sharing with the firms, nor does it take a percentage of revenue in our setup. The exogenous nature of λ is motivated by two key factors: (1) simplification for tractability, which helps focus on the impact of personalized versus unpersonalized rank-

ings on firm pricing strategies, and (2) real-world relevance, as many platforms implement ranking algorithms based on fixed, pre-determined criteria (at least within a sufficiently long period of time), which do not dynamically adjust to firm-specific pricing strategies.

A.4 Implementation

A.4.1 Action space

We discretize the action space to meet the requirements of Q-learning, and compute the lowest numerical Nash-Bertrand equilibrium prices p_j^N and highest monopoly prices p_j^M for each firm j that maximizes the joint profits for one shot game, for each parameter combination. For pure-strategy Nash-Bertrand equilibrium, we use the unique set of prices. Without pure-strategy Nash-Bertrand equilibrium, we use the lowest price in the mixed-strategy Nash-Bertrand equilibrium. To construct the action space for firm j, we utilize the values of p_j^N and p_j^M. The action space for each firm j is defined as $[p_j^N - \xi(p_j^M - p_j^N), p_j^N + \xi(p_j^M - p_j^N)]$ with equal step size 0.25. Here, ξ is a parameter such that the feasible action space ranges from below competitive prices to above monopoly prices.

A.4.2 State Space

In the context of reinforcement learning, the state space refers to the information that an algorithm can use to determine its actions. In our study, we constrain the state space to only include information from the preceding period. Specifically, the state is represented as a set comprising of the seller prices in the previous period:

$$s_t = \{p_j^{t-1}\}_{j=1}^N \tag{A19}$$

This specific state construction confers several benefits. Due to this construction, an agent's pricing strategy becomes dependent on the prior prices of its competitors. With access to this information, the agent can learn from the pricing tactics of its adversaries and adjust its own behavior accordingly when those strategies change over time. Additionally, by linking current prices to past prices, agents can detect deviations from cooperative

behavior and employ punishment strategies to promote long-term cooperation in repeated price competition scenarios.

A.4.3 Exploration

In the context of reinforcement learning, agents are faced with a challenging exploration-exploitation trade-off, whereby they must decide whether to exploit a learned policy to maximize immediate rewards or to explore new actions to acquire information and potentially discover more profitable actions in the future. To address this dilemma, various approaches have been proposed in the literature. In this study, we adopt the time-declining $\epsilon-$greedy approach to balance the exploration-exploitation trade-off. Specifically, we set the exploration rate, ϵ_t, to be a function of time that decreases over time, formulated as:

$$\epsilon_t = e^{-\beta t} \tag{A20}$$

where $\beta > 0$ is a parameter. At the outset, the agent selects actions at random, but as time progresses, it favors actions that yield higher returns more frequently. A larger value of β results in a faster decay of the exploration rate. This approach strikes a balance between exploiting known information and exploring new possibilities, allowing for a gradual shift towards exploiting learned policies as the agent gains experience.

A.4.4 Initialization

The initialization of Q-values plays a crucial role in the Q-learning algorithm, as it can significantly impact the quality of the learned policy and the learning process. A common approach to Q-value initialization is to incorporate domain-specific knowledge, which can speed up learning and improve the quality of the learned policy. In our specific setup, the agents use the time-declining $\epsilon-$greedy approach to choose actions, and as such, the agent's initial prices tend to be exploratory or random. Therefore, it is appropriate to initialize the Q-values based on the expected discounted rewards that would occur if the competing agents set prices randomly. To this end, we set the initial Q-values, Q_0,

accordingly:

$$Q_0^j(s, p_j) = \frac{\sum_{p_{-j} \in \mathcal{P}_{-j}} R_j(p_j, p_{-j})}{(1-\delta)|\mathcal{P}_{-j}|} \tag{A21}$$

A.4.5 Updating Rule

The updating rule describes how $Q^j(s_t, a)$ is updated at each iteration. The value matrix Q^j is updated as follows

$$Q^j(s_t, p) \leftarrow (1-\alpha)Q^j(s_t, p) + \alpha(R_t^j + \delta \max_{p'} Q^j(s_{t+1}, p')) \tag{A22}$$

where α is the learning rate that captures how much new information overrides the existing information in the Q-values.

A.4.6 Convergence

The convergence of multi-agent independent Q-learning algorithms cannot be guaranteed. Therefore, following Calvano et al. (2020), we utilize a convergence verification rule to determine when stable behavior is achieved. Specifically, we consider the algorithm to have achieved convergence if the optimal action given any state does not change for 100000 consecutive periods. Let $p_j^t(s) = \operatorname{argmax}_p Q_t^j(s, p)$ denote the optimal action for agent j at time t and state s. Then, we require that $p_j^t(s) = p_j^{t+1}(s)$ holds for all $t = T-100000, \ldots, T-1$, where T is the total number of periods. Once convergence is verified, the learning process is stopped and the pricing algorithms have reached their rest points. In our experiment, we allow the algorithm to run for as many rounds as necessary for convergence, with the number of steps dependent on factors such as the type of ranking systems, number of products, exploration rate, and learning rate.

A.5 Optimality of Multi-agent Reinforcement Learning algorithms

In a stationary environment, the Reinforcement learning algorithms can reach optimal prices Watkins (1989). However, when the environment is non-stationary, such as the case that we have where multiple sellers are using RL algorithms, the RL algorithms are not guaranteed to reach optimal prices. Recall that the price competition in our setting

is a repeated game. We know from Folk theorem that any prices between Bertrand Nash prices and the monopoly prices (perfect collusion) can be sustained in a repeated game equilibrium. We find that the RL algorithms in our case are able to converge to prices that are higher than the Bertrand Nash Prices but lower than the Monopoly prices. Therefore we focus on characterizing the resting point equilibrium as in Calvano et al. (2020), Brown and MacKay (2023), and Johnson et al. (2023) instead of examining the learning path to reach the resting point as in Hansen et al. (2021). The main objective of the paper is to study how personalization in product ranking affects the *converged* prices and the resulting consumer welfare when algorithms are delegated to make pricing decisions.

A.6 Details of Multi-Agent UCB-tuned Algorithm

Algorithm A.1 Multi-Agent UCB-tuned

Initialize T, K
$t \leftarrow 1$
for $j = 1$ to n **do**
  $n_{jk}^t \leftarrow 0$
end for
while $t < T$ **do**
  for $j = 1$ to n **do**
    $$p_{jk}^t = \begin{cases} \text{Randomly choose each action once} & \text{if } t \leq K \\ \text{argmax}_k \text{ UCB-tuned}_{jk}^t & \text{if } t > K \end{cases}$$
    $n_{jk}^t \leftarrow n_{jk}^t + 1$
  end for
  Execute actions $p = (p_{1k_1}^t, \ldots, p_{nk_n}^t)$ and observe reward R_{jk}^t
  for j = 1 to n **do**
    $V_{jk}^t = \overline{(R_{jk}^t)^2} - \overline{R_{jk}^t}^2 + \sqrt{\frac{2 \log t}{n_{jk}^t}}$
    $\text{UCB-tuned}_{jk}^t = \overline{R_j^t} + \sqrt{\frac{\log t}{n_{jk}^t} min(\frac{1}{4}, V_{kt})}$
  end for
  $t \leftarrow t + 1$
end while

A.7 Risk-averse Consumers

We further consider the possibility that consumers are risk-averse. To account for consumers' risk preferences, following Erdem and Keane (1996), we modify the utility func-

tion in Equation (1) to

$$u_{ij} = \tilde{a}_{ij} - \phi p_j - r\tilde{a}_{ij}^2, \forall j = 0, 1, \ldots, n. \text{ where } \tilde{a}_{ij} = a_j + \mu\epsilon_{ij} \tag{A23}$$

where r is the risk coefficient. Consumers are risk averse when $r > 0$. The utility is still linear in price but concave in a_j and ϵ_{ij} for $r > 0$. This utility formulation captures the possibility that consumers prefer to choose products with less uncertainty.

The corresponding expected utility for product j before search is

$$E[u_{ij}] = E[\tilde{a}_{ij}] - \phi p_j - rE[\tilde{a}_{ij}^2] = a_j + \mu\gamma - r(a_j + \mu\gamma)^2 - \phi p_j - r\frac{\pi^2}{6} \tag{A24}$$

We compare the results under perfect personalized ranking and unpersonalized ranking. Table A.1 presents the results for risk-averse consumers, showing that equilibrium prices derived from Q-learning algorithms are lower under both unpersonalized and personalized ranking systems compared to equilibrium prices under the risk-neutral consumer assumption. This outcome can be explained by the reduced search gains (or value of continuing to search) associated with lower-ranked products for risk-averse consumers, who are therefore less inclined to explore beyond the top-ranked options. Firms, expecting this, have stronger incentives to lower prices to secure better rankings, as price is their primary lever for influencing rank position. Consequently, profits also decline when consumers are risk averse under both ranking systems. Consumer surplus is smaller due to concave utility with respect to product quality and their risk aversion. Nonetheless, most importantly, our main result remains robust: even when consumers are risk averse, unpersonalized ranking leads to lower prices and profits, while generating higher consumer surplus compared to personalized ranking.

Table A.1: Results when consumers are risk averse

	Unpersonalized	Personalized	% Change	KS	MW
Q-learning price	2.57 (0.23)	3.1 (0.12)	+20%	0.94***	9798.0***
Profit	0.43 (0.07)	0.71 (0.02)	+62%	1.0***	10000.0***
Consumer surplus	2.0 (0.2)	1.83 (0.09)	-8%	0.65***	2385.0***

Note: The parameter set is: $a_0 = 0.0, a_1 = 4.0, a_2 = 4.5, \rho = 1.5, \phi = 1.0, \mu = 1.0, mc_1 = 1.0, mc_2 = 1.5,\ n = 2, \lambda = 1.0, \gamma = -1.0, \alpha = 0.1, \beta = 2e-06$. All variables are averaged over the last 1000 steps. KS indicates the statistics for the Kolmogorov-Smirnov test. MW indicates the statistics for the Mann-Whitney U rank test.

References

Brown, Z. Y. and MacKay, A. (2023). Competition in pricing algorithms. American Economic Journal: Microeconomics, 15(2):109–156.

Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. American Economic Review, 110(10):3267–97.

Erdem, T. and Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. Marketing science, 15(1):1–20.

Hansen, K. T., Misra, K., and Pai, M. M. (2021). Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. Marketing Science, 40(1):1–12.

Johnson, J. P., Rhodes, A., and Wildenbeest, M. (2023). Platform design when sellers use pricing algorithms. Econometrica, 91(5):1841–1879.

Watkins, C. J. C. H. (1989). Learning from delayed rewards.