«Московский физико-технический институт (национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Гунаев Руслан Гуламович

Онлайн ценообразование с помощью структурированных многоруких бандитов

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа Бакалавра

Научный руководитель:
Дорн Юрий Викторович

Москва
2021

Содержание

Аннотация

Большинство онлайн-рынков характеризуются конкурентной средой. Из-за сложности таких рынков трудно разработать эффективные стратегии ценообразования. В данной работе предложен алгоритм WEIGHTED UCB+QBC для решения задачи ценообразования в страховании. Предложенный алгоритм является усовершенствованием алгоритма UCB. Модификация заключатеся в использовании активного обучения, которое позволяет использовать различные параметризации неизвестной функции спроса, делая наиболее правильный отбор точек в процессе работы алгоритма. Это позволяет сокращать траты во время проведения экспериментов. Также мы используем весовые функции, позволяющие отдавать предпочтение только хорошим параметризациям, также учитывать риски использования цен, сильно удаленных от уже проверенных. Главная цель алгоритма – как можно быстрее находить оптимальную цену для продажи страховки. В результате экспериментов на данных компании Тинькофф мы получили увеличение прибыли на 20% по сравнению с политикой фиксированных цен.

1 Введение

2 Постановка задачи

2.1 Задача динамического ценообразования

Формально задачу динамического ценообразования можно поставить следующим образом: требуется найти последовательность цен $\hat{\mathbf{X}}(T) = (x_1, x_2, \ldots, x_T)$ такую, что прибыль $r(x)$ будет максимальна. Существует 4 варианта постановки:

1. $$x^* = \arg\max_{x(T)} \mathsf{E}[r(x(T))].$$

   В этой постановке нам необходимо найти оптимальную цену в конкретный момент времени T.

2. $$x^* = \arg\max_{x(t)} \mathsf{E}[r(x(t))].$$

   Необходимо найти оптимальную цену за наименьшее время.

3. $$\hat{\mathbf{X}}(T) = \arg\max_{\mathbf{X}(T)} \sum_{t=1}^{T} \mathsf{E}[r(x_t)].$$

   Здесь мы хотим оптимизировать всю траекторию цен.

4. Найти зависимость $r(x)$.

Первые две постановки нам не подходят, потому что во время тестирования алгоритма мы будем тестировать неоптимальные цены, тем самым теряя деньги компании, поэтому в рамках данной работы мы сконцентрируемся на 3 постановке.

Также необходимо учитывать бизнес-ограничения.

1. Нельзя менять цену слишком резко $\dfrac{x_{i+1}}{x_i} \sim 1$ для любого момента времени i.

2. Экспертные ограничения на цену $x_{\min} \leqslant x_i \leqslant x_{\max}$ для любого момента времени i.

3. Каждый эксперимент стоит денег.

Таким образом, можем записать итоговую постановку задачи. Найти $\hat{\mathbf{X}}(T)$ такую, что

$$\hat{\mathbf{X}}(T) = \arg\max_{\mathbf{X}(T)} \sum_{t=1}^{T} \mathsf{E}[r(x_t)] \\ s.t\ x_{\min} \leqslant x_t \leqslant x_{\max}\ \forall t \leqslant T .$$

В текущей постановке выполнены все три ограничения: в процессе эксперимента цена подбирается так, чтобы максимизировать общую прибыль, оптимальная цена лежит в нужном диапазоне, внутри которого разброс цен небольшой.

2.2 Функция спроса

Теперь добавим специфику. Нашу целевую функцию прибыли можно представить в виде

$$r(x) = Q(x) \cdot x,$$

где $Q(x)$ – число проданных страховок по цене x, иначе говоря спрос. Проблема заключается в том, что мы не знаем настоящую зависимость спроса от цены, поэтому предлагается использовать различные ее параметризации:

1. линейная функция: $Q(x) = \max\{-ax + b, 0\}$;
2. гиперболическая функция: $Q(x) = \max\{-\frac{a}{x} + b, 0\}$;
3. экспоненциальная функция: $Q(x) = \max\{-\exp(ax + b)c + d, 0\}$;
4. показательная функция: $Q(x) = \max\{ba^x + c, 0\}$.

Эти параметризации не обязаны точно аппроксимировать функцию спроса, основная задача заключается в том, чтобы в результате работы алгоритма обновлять параметры так, чтобы аппроксимации верно указывали на оптимальную цену.

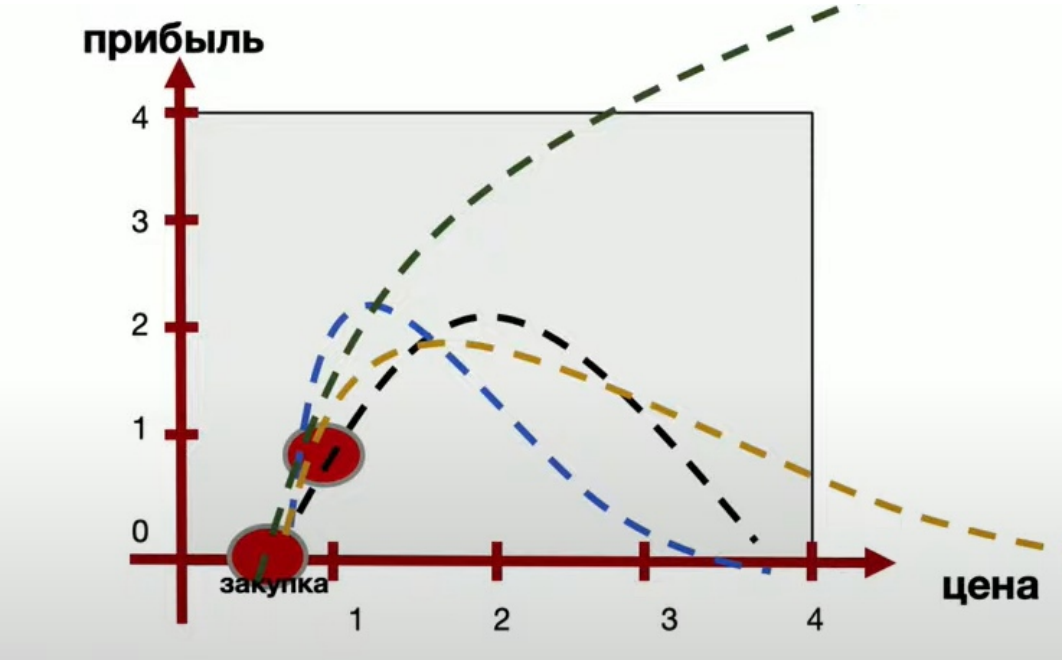

Рис. 1: Пример аппроксимации по двум точкам.

3 Многорукие бандиты

3.1 Алгоритм UCB

Положим $n_{i,t}$ – количество раз, когда была сыграна ручка i до момента времени t. r_t – награда, которую мы получаем в момент времени t. $I_t \in \{1, 2, \ldots, N\}$ – выбранная ручка в момент времени t. Эмпирическая оценка награды ручки i в момент t:

$$\hat{\mu}_{i,t} = \frac{\sum_{s=1:I_s=i}^{l} r_s}{n_{i,t}}.$$

Регрет задается следующим образом

$$R(T) = T\mu^\star - \sum_{t=0}^{T-1} \mathsf{E}\left[r^t\right] \tag{1}$$

UCB присваивает каждой ручке в каждый момент времени следующее значение:

$$\mathrm{UCB}_{i,t} := \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}}$$

Algorithm 1: UCB algorithm

Data: N arms, number of rounds $T \geq N$
for $t = 1 \ldots N$ **do**
  play arm t
end
for $t = N + 1 \ldots T$ **do**
  play arm
$$I_t = \arg\max_{i \in \{1 \ldots N\}} \mathrm{UCB}_{i,t-1}$$
end

Theorem 3.1 (Верхняя оценка ожидаемого регрета UCB алгоритма)**.** *Пусть $R(T)$ – регрет UCB алгоритма для некоторого многорукого бандита, тогда для любого T верна верхняя оценка*

$$\mathsf{E}[R(T, \Theta)] \leq \sum_{i:\mu_i<\mu^*} \frac{4 \ln T}{\Delta_i} + 8\Delta_i, \;\; \Delta_i = \mu^* - \mu_i.$$

Доказательство. Есть более фундаментальная причина выбора $\sqrt{\frac{\ln t}{n_{i,t}}}$. Эта верхняя оценка вытекает из неравенства Чернова-Хоффдинга. Для каждой ручки верно

$$|\hat{\mu}_{i,t} - \mu_i| < \sqrt{\frac{\ln t}{n_{i,t}}}$$

с вероятностью не меньше $1 - 2/t^2$. Их этого получаем два важных неравенства:

1. Нижняя граница для $\mathbf{UCB}_{i,t}$. С вероятностью не меньше $1 - 2/t^2$,

$$\text{UCB}_{i,t} > \mu_i$$

2. Верхняя граница для $\hat{\mu}_{i,t}$ с большим числом семплов. При $n_{i,t} \geq \frac{4 \ln t}{\Delta_i^2}$, с вероятностью не меньшей $1 - 2/t^2$ верно,

$$\hat{\mu}_{i,t} < \mu_i + \frac{\Delta_i}{2}$$

1 показывает, что значение UCB, вероятно, равно истинному вознаграждению: в этом смысле алгоритм UCB оптимистичен. 2 – что при наличии достаточного количества (а именно, по крайней мере, $\frac{4 \ln t}{\Delta_i^2}$) семплов оценка вознаграждения, вероятно, не превышает истинное вознаграждение более чем на $\Delta_i/2$. Эти ограничения показывают, что алгоритм быстро находит субоптимальную ручку.

Lemma 3.2. *В любой момент времени t, если субоптимальная ручка i ($\mu_i < \mu^*$) была сыграна $n_{i,l} \geq \frac{4 \ln t}{\Delta_i^2}$ раз, тогда $UCB_{i,t} < UCB_{I^*,t}$ с вероятностью $1 - 4/t^2$. Это значит, что любого t,*

$$P\left(I_{t+1} = i \mid n_{i,t} \geq \frac{4 \ln t}{\Delta_i^2}\right) \leq \frac{4}{t^2}$$

Доказательство.

$$\begin{aligned}
\text{UCB}_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}} &\leq \hat{\mu}_{i,t} + \frac{\Delta_i}{2} && \text{при } n_{i,t} \geq \frac{4 \ln L}{\Delta_i^2} \\
&< \left(\mu_i + \frac{\Delta_i}{2}\right) + \frac{\Delta_i}{2} \\
&= \mu^* && \text{при } \Delta_i := \mu^* - \mu_i \\
&< \text{UCB}_{i^*,t}
\end{aligned}$$

□

Lemma 3.3. *Пусть $n_{i,T}$ – количество раз, когда ручка i была выбрана алгоритмом . Тогда для любой ручки с $\mu_i < \mu^*$,*

$$\mathsf{E}\left[n_{i,T}\right] \leq \frac{4\ln T}{\Delta_i} + 8$$

Доказательство. Для любой ручки i ожидаемое число раз, когда она была сыграна

$$\begin{aligned}
\mathsf{E}\left[n_{i,T}\right] &= 1 + \mathsf{E}\left[\sum_{t=N}^{T} \mathsf{I}\left(I_{t+1} = i\right)\right] \\
&= 1 + \mathsf{E}\left[\sum_{t=N}^{T} \mathsf{I}\left(I_{t+1} = i, n_{i,t} < \frac{4\ln t}{\Delta_i^2}\right)\right] + \mathsf{E}\left[\sum_{t=N}^{T} \mathsf{I}\left(I_{t+1} = i, n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right)\right. \\
&\leq \frac{4\ln T}{\Delta_i^2} + \mathsf{E}\left[\sum_{t=N}^{T} \mathsf{I}\left(I_{l+1} = i, n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right)\right] \\
&= \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} P\left(I_{l+1} = i, n_{i,l} \geq \frac{4\ln t}{\Delta_i^2}\right) \\
&= \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} P\left(I_{t+1} = i \mid n_{i,t} \geq \frac{4\ln T}{\Delta_i^2}\right) P\left(n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right) \\
&\leq \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} \frac{4}{t^2} \\
&\leq \frac{4\ln T}{\Delta_i^2} + 8
\end{aligned}$$

□

Тогда пользуясь леммами, итоговый ожидаемый регрет до времени T:

$$\mathsf{E}[R(T,\Theta)] = \sum_{i:\mu_i<\mu^*} \mathsf{E}\left[n_{i,T}\right]\Delta_i \leq \sum_{i:\mu_i<\mu^*} \frac{4\ln T}{\Delta_i} + 8\Delta_i$$

□

3.2 Активное обучение

В силу того, что во время проведения эксперимента мы не можем рисковать, проверяя плохие цены, в данной работе предлагается использовать

активное обучение. Цель активного обучения заключается в том, чтобы достичь как можно лучшего качества, использовав при этом как можно меньше примеров.

3.3 UCB+QBC

3.4 WEIGHTED UCB+QBC

4 Вычислительный эксперимент

5 Заключение