# Оглавление

[Оглавление 2](#_Toc148925163)

[Глава 1. Теория 3](#_Toc148925164)

[Общие понятия обучения с подкреплением 3](#_Toc148925165)

[Алгоритмы обучения с подкреплением (общий случай) 4](#_Toc148925166)

[Задача о многоруком бандите. Алгоритмы решения 5](#_Toc148925167)

# Глава 1. Теория

## Общие понятия обучения с подкреплением

Обучение с подкреплением (Reinforcement Learning, RL) — это область машинного обучения, где агент обучается взаимодействовать со средой для максимизации некой награды. В отличие от обучения с учителем, в RL у агента нет явных примеров правильного поведения или сведений об окружающей среде, и он определяет оптимальную стратегию взаимодействия со средой на основе полученных наград. Такой метод обучения наиболее близок к человеческому, в котором не приходится заранее собирать и анализировать большое количество информации (обучающую выборку) для построения прогнозов. Вместо этого человек выбирает как ему лучше поступать, основываясь на собственном ранее полученном опыте.

В последние годы обучение с подкреплением сыскало большую популярность в виду универсальности своего математического аппарата. Его методы нашли применение в задача следующих прикладных задачах:

* Рекомендация новостных статей пользователям;
* Показ рекламы в Интернете;
* Управление технологическими процессами;
* Управление роботами;
* Управление ценами и ассортиментом в сетях продаж;
* Игра на бирже;
* Маршрутизация в телекоммуникационных сетях;
* Маршрутизация в беспроводных сенсорных сетях;
* Стратегические игры: шахматы, го, Dota2, StarCraft2…

Основными компонентами обучения с подкреплением являются:

* Агент: сущность, принимающая решения.
* Среда: всё, что окружает агента и на что он может воздействовать.
* Действие: взаимодействие агента со средой.
* Состояние: описание среды в данный момент времени.
* Награда: сигнал, который агент получает от среды после выполнения действия в определенном состоянии.
* Оптимальная стратегия: совокупность поочередных действий агента, приводящих к наибольшему количеству благоприятных наград.
* Раунд: интервал, в которой происходит взаимодействие агента со средой.

## Алгоритмы обучения с подкреплением (общий случай)

Алгоритмы обучения с подкреплением охватывают широкий спектр методов, которые агенты используют для обучения оптимальной стратегии взаимодействия с окружающей средой. Основная цель агента — максимизировать кумулятивную награду в течение времени.

Эти алгоритмы можно классифицировать в зависимости от того, используют ли они модель среды в процессе обучения:

1. **Основанные на модели**

В этих алгоритмах агент пытается узнать или аппроксимировать модель среды.

* Динамическое программирование (Dynamic Programming): это классический метод, который использует знание модели для нахождения оптимальной политики. Примеры включают в себя итерацию по политике (Policy Iteration) и итерацию по функции ценности (Value Iteration).
* Планирование (Planning): агент использует свою модель среды для имитации возможных будущих сценариев и принятия решений на основе этих сценариев. Например, Monte Carlo Tree Search (MCTS) использует случайные сэмплы для исследования пространства действий.
* Адаптивное планирование: при изменении среды агент может адаптировать свою модель и соответственно изменять свою политику. Примером такого алгоритма является метод временных разностей (Temporal-Difference, TD).

1. **Без модели**

Агент не пытается явно узнать модель среды, но вместо этого напрямую оценивает ценность состояний или действий.

* Q-learning: оффлайн метод, который оценивает функцию ценности действия, не зависящую от текущей политики агента.
* SARSA: оценивает функцию ценности действия, основываясь на текущем и следующем действиях.
* Policy Gradient Methods: вместо оценки функции ценности агент напрямую оптимизирует свою политику с использованием градиентного восхождения.
* Actor-Critic Methods: комбинируют идеи из методов, основанных на политике, и методов оценки ценности, имея две компоненты: актер, выбирающий действия, и критик, оценивающий действия.

**Вывод**

Выбор между алгоритмами, основанными на модели, и без модели зависит от специфики задачи, доступности данных, вычислительных ресурсов и других факторов. В то время как методы, основанные на модели, могут быть более эффективными в терминах количества требуемых взаимодействий со средой, поскольку они могут "думать вперед", методы без модели часто проще в реализации и могут быть более устойчивыми в сложных или непредсказуемых средах.

## Задача о многоруком бандите. Алгоритмы решения

До текущего момента обучение с подкреплением рассматривалось как задача о среде с состояниями. То есть среда могла меняться и адаптироваться, реагируя на действия агента, что оказывало влияние на дальнейший выбор действий. Однако существует частный случай обучения с подкреплением, в котором среда является стационарной. Такая задача получила название задачи о многоруком бандите (Multi-Armed Bandit).

**Постановка задачи:**

Пусть – множество всех возможных действий агента.

– неизвестное распределение премии для действия .

– стратегия агента в момент , распределенная на .

Тогда игра агента со средой производится следующим образом: