Разработка алгоритма на основе методов динамического программирования.

1. Введение в динамическое программирование.

Динамическое программирование – метод решения сложных задач, в котором вместо одной сложной задачи решаются несколько более простых подзадач, на которые она разбита, после чего для каждой подзадачи сохраняются решения [3].

В обучении с подкреплением под динамическим программированием понимают семейство алгоритмов, которые используются для вычисления оптимальных стратегий, при условии, что имеется модель окружающей среды в виде марковского процесса принятия решений [1].

Ключевая идея ДП и ОП вообще заключается в использовании функций ценности для организации и структурирования поиска хороших стратегий. Стратегию можно представить функцией $\pi(s): S \to A$, указывающей какое действие должно выполняться в каждом состоянии. Конечной целью является нахождение оптимальной стратегии, задающей для каждого состояния действие, максимизирующее конечный доход. Функция ценности состояния указывает насколько хорошо для агента пребывание в конкретном состоянии при стратегии π, то есть оценивает ценность состояния при следовании стратегии и определяется как:

$$\begin{aligned} v_\pi(\mathrm{s}) &= E_\pi\left[G_t \vee S_t = s\right] ¿ E_\pi\left[R_{t+1} + y * G_{t+1} \middle| S_t = s\right] \\ ¿ &\sum_a \pi(a|s) \sum_{\acute{s}} \sum_r p(\acute{s}, r|s, a)\left[r + y * E_p\left[G_{t+1} \vee S_{t+1} = \acute{s}\right]\right] \\ ¿ &\sum_a \pi(a|s) \sum_{s',r} p(\acute{s}, r|s, a)\left[r + y * v_\pi(\acute{s})\right] для всех\ s \in S \end{aligned} \quad (1)$$

где действия a берутся из множества A(s), следующие состояния $\acute{s}$ берутся из множества S, в которые переходит агент после совершенного действия, и вознаграждения r берутся из множества R. Равенство (1) называется уравнением Беллмана для функции ценности состояния s при стратегии π, обозначаемая v_π[1].

При этом для детерминированной стратегии и с учетом ограниченного количества шагов получим

$$v_\pi(\mathrm{s}) = \sum_{s',r} p(\acute{s}, r|s, a)\left[r + y * v_\pi(\acute{s})\right] \quad (2)$$

Данное равенство представляет собой систему линейных уравнений, которую можно представить следующим образом

$$\begin{aligned} v_{\pi}(s_n) &= r_n + y * v_{\pi}(s_{n-1}) \\ v_{\pi}(s_{n-1}) &= r_{n-1} + y * v_{\pi}(s_{n-2}) \\ &\dots \\ v_{\pi}(s_1) &= r_1 \end{aligned} \qquad (3)$$

Однако для решения системы уравнений (3) требуется прямолинейное длительное вычисление с использованием большого объема памяти, экспоненциально растущего в зависимости от n. В этой связи для решения уравнения Беллмана [1] разработаны итеративные методы:

- итерация по стратегиям;
- итерация по ценности.

1.2 Решение задачи поиска оптимальной стратегии итерацией по стратегиям.

При итерации по стратегиям алгоритм начинается с инициализации ценности всех состояний S случайной величиной (например 0), после чего инициализируется случайная стратегия и проводится расчет ценности по формуле 2. Далее необходимо проверить функцию V(s) на оптимальность, и если функция не оптимальна, провести улучшение стратегии с последующим итерационным вычислением функции ценности. Типовой алгоритм итерации по стратегиям представлен на рисунке 2.

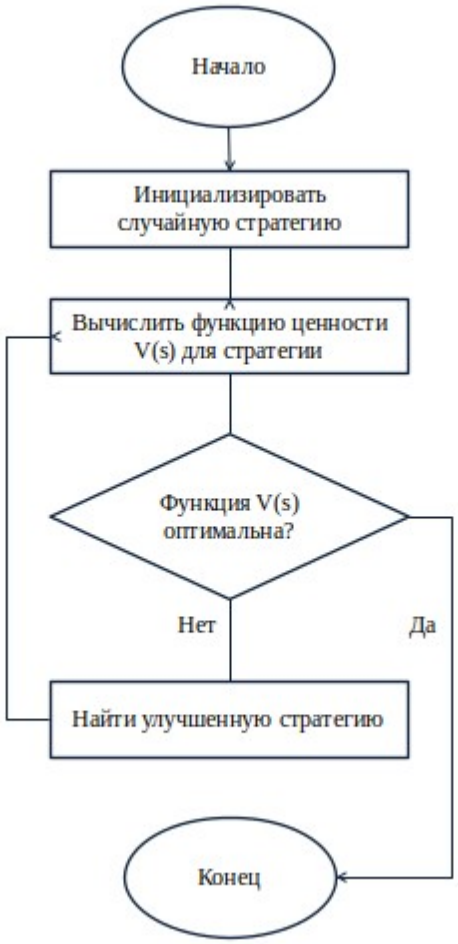


Рисунок 2.

Функция ценности рассматривается как функция состояние/действие и имеет название Q-функция. Она показывает, насколько хорошо для агента выполнять конкретное действие в соответствии со стратегией π и определяется

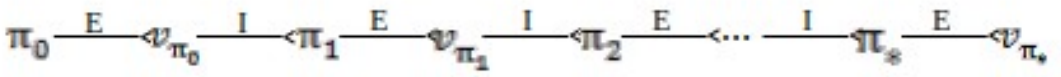
$$q_\pi(\text{s},\text{a}) = E_\pi\left[R_{t+1} + y * v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right] ¿ \sum_{s',r} p(\acute{s}, r | s, a)\left[r + y * v_\pi(\acute{s})\right] \quad (4)$$

Q-функция играет важную роль в улучшении стратегии, поскольку если существует такая стратегия, что выполняется следующее неравенство $q_\pi(\text{s}, a) > v_\pi(\text{s}))$ для одного состоя s, при условии, что в остальном стратегия $\acute{\pi}$ соответствует стратегии π, тогда стратегия $\acute{\pi}$лучшестратегии π. Тогда, при рассмотрении изменения всех возможных действий во всех состояниях, выбирая в каждом состоянии то действие, которое максимизирует Q-функцию, получим новую жадную стратегию $\acute{\pi}$, определенную следующим образом:

$$\acute{\pi}(s) = argmax_a q_\pi(\text{s ,a}) ¿ argmax_a E\left[R_{t+1} + y * v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right]$$
$$¿ argmax_a \sum_{s',r} p(\acute{s}, r|s, a)\left[r + y * v_\pi(\acute{s})\right] \quad (5)$$

Жадная стратегия выбирает действие, которое является наилучшим в краткосрочной перспективе согласно функции v_π. Процесс конструирования новой стратегии, улучшающий исходную путем жадного выбора относительно функции ценности исходной стратегии, называется улучшением стратегии.

Итерационное улучшение стратегии необходимо проводить до получения оптимальной стратегии – стратегии максимизирующей вероятный доход. Критерий оптимальности стратегии определяется равенством новой, выработанной на очередном шаге, жадной стратегии $\acute{\pi}$ и стратегии предыдущего шага π, поскольку тогда $v_\pi = v_{\acute{\pi}}$ и из формулы 5 следует, что для всех $s \in S$. Таким образом, итерационное улучшение стратегии до оптимальной можно представить следующей схемой:



где E – это оценивание стратегии, а I – улучшение стратегии.

При этом, в соответствии с вышеизложенным, каждая следующая стратегия лучше предыдущей, только если предыдущая уже не является оптимальной. Для конечного МППР существует конечное число стратегий, поэтому процесс должен сойтись к оптимальной стратегии и оптимальной функции ценности за конечное число итераций. Данный способ нахождений оптимальной стратегии и является способом итерации по стратегиям. Таким образом, полный алгоритм итерации по стратегиям будет выглядеть следующим образом:

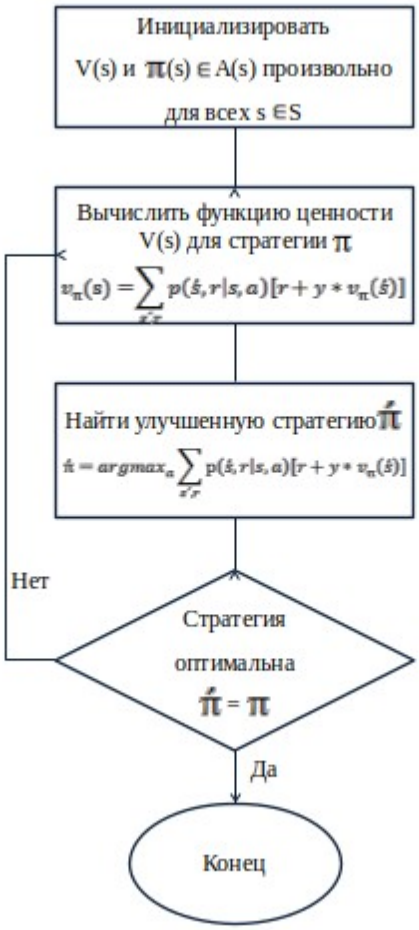


Рисунок 3.

Для решения задачи построения траектории облета зон противодействия сил ПВО противника реализовав данный алгоритм в программном обеспечении и расчитав ценности каждого состояния с учетом текущего расположения ПВО получим:

вставить скриншот ПО

Теперь для построения маршрута облета зон ПВО из любой точки необходимо, всего лишь, построить маршрут по точкам (состояниям) жадно (глубина расчета 1 шаг) выбирая действие максимизирующее значение Q-функции.Выбрав за начальную точку агента s(2, 2), объекты ПВО s(x,x) и s(x,x) и объект назначения s(x,x) получим:

вставить скриншот ПО

Задача построения маршрута облета зон ПВО методом динамического программирования с использованием итерации по стратегиям решена.

1.3 Решение задачи поиска оптимальной стратегии итерацией по ценности.

При итерации по ценности мы сначала инициализируем функцию ценности некоторым случайным значением, после чего перебираем все состояния и находим новую функцию ценности

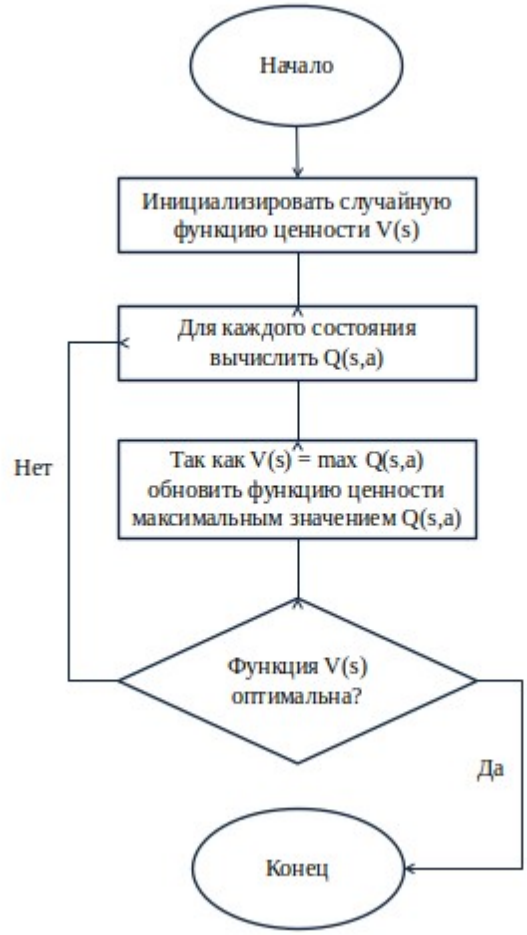


Рисунок 3.

Итерация по ценности – оптимизация происходит без учета стратегии, алгоритм следующий: инициализируем ценность состояний случайными числами, например $V(s)=0$ и перебором смотрим все возможные значения Q-функции (функции ценности пары состояние/действие) с учетом текущего вознаграждения за действие $a \in A$ и дисконтированной ценности следующего состояния $Q(S,A)=R_{S\dot{S}}^{A}+y*V(\dot{S})$ при этом ценность текущего состояния определяем как максимальное из возможных результатов Q-функции $V(s)=argmax_a Q(S,A)$. Поскольку сходимость возможна только в пределе, но стратегии становится оптимальной раньше, вводим Ɛ– маленькое положительное число, и при изменении ценности всех состояний на значение меньше Ɛостанавливаем итерации. При этом стратегия максимизирующая $V(s)$ будет являться оптимальной.