Лекция 4: Градиент стратегии и модель актор-критика

Панов А.И.
panov.ai@mipt.ru

Центр когнитивного моделирования МФТИ
Интситут проблем искусственного интеллекта ФИЦ ИУ РАН

2020 – Машинное обучение с подкреплением
Курс для Сбербанка

План лекции

1. Прямой поиск стратегии
2. Теорема о градиенте стратегии
3. Монте-Карло градиент стратегии
4. Базовый уровень
5. Метод базового уровня
6. Оценка стратегии с помощью критика
7. Алгоритм A3C
8. Алгоритм DDPG

Прямой поиск стратегии в RL

- Ранее мы использовали аппроксимацию функции полезности состояния или действия, параметризованную с помощью w:

$$\hat{V}(s, \mathrm{w}) \approx V^{\pi}(s),$$

$$\hat{Q}(s, a, \mathrm{w}) \approx Q^{\pi}(s, a)$$

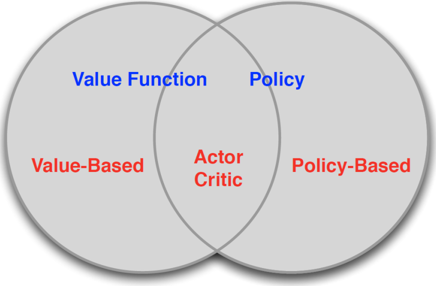
- Стратегия генерировалась напрямую по функции полезности (например, ϵ-жадно)
- Однако, можно напрямую параметризовать стратегию:

$$\hat{\pi}(s, \theta) = \mathbb{P}[a|s, \theta]$$

- Вапник (1998) – не нужно решать общую задачу через промежуточные шаги

Типология методов RL

- Основанные на полезности (value-based):
  - легко интерпретируемы,
  - могут концентрироваться на не самых важных признаках
- Поиск стратегии (policy-based):
  - цель поиска – сама стратегия,
  - игнорируются другие полезные данные
- Актор-критик (actor-critic)
  - строят как функцию полезности,
  - так и стратегию



Преимущества методов, основанных на стратегии

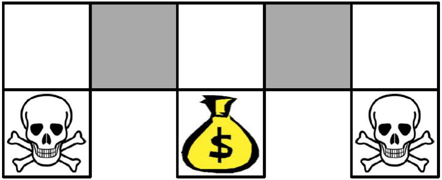
Плюсы:

- Лучшие свойства сходимости процесса обучения
- Эффективны в задачах большой размерности и непрерывных пространствах действий
- Могут обучаться стохастическим стратегиям
- Иногда стратегии проще, чем функции полезности

Минусы:

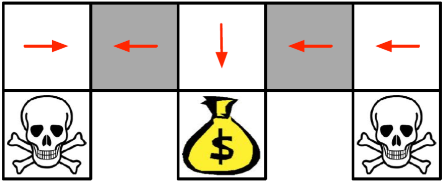
- Обычно сходятся к локальному оптимуму, а не к глобальному (особенно с нелинейными аппроксиматорами)
- Полученная модель обычно специфична для конкретной задачи и плохо обобщаема
- Обычно процесс вычисления стратегии неэффективен и обладает высокой дисперсией

Стохастическая стратегия: затемненный клеточный мир



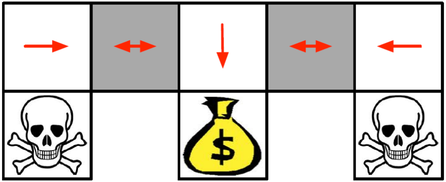
- Агент не различает серые клетки
- Рассмотрим признаки следующего вида (для всех направлений N, E, S, W):
$$\phi(s,a) = (1\ 0\ 1\ 0|0\ 1\ 0\ 0)$$
стена к северу и югу, идем на восток
- Сравним метод, основанный на полезности с аппроксимацией функции:
$$\hat{Q}(s,a,\mathrm{w}) = f(\phi(s,a),\mathrm{w}),$$
- с методом, основанным на стратегии с параметризацией:
$$\hat{\pi}(s,a,\mathrm{w}) = g(\phi(s,a),\mathrm{w})$$

Пример: альтернативный клеточный мир



- Оптимальная **детерминированная** стратегия будет следующей:
  - двигаться на W в обоих серых состояниях (красные стрелки),
  - двигаться на E в обоих серых клетках
- В любом случае, агент может застрять и никогда не отыскать целевого состояния
- Основанные на полезности методы находят близкую к детерминированной стратегию (ϵ-жадную)
- В этом случае агент может блуждать по коридору длительное время

Пример: альтернативный клеточный мир



- Оптимальная стохастическая стратегия состоит в том, чтобы двигаться случайно на E или на W в серых клетках:

$$\hat{\pi}(\text{стена на N или S}, \text{двигаться на E}, \mathrm{w}) = 0.5,$$

$$\hat{\pi}(\text{стена на N или S}, \text{двигаться на W}, \mathrm{w}) = 0.5$$

- Это позволит достичь целевого состояния за небольшой количество шагов с высокой вероятностью
- Основанные на стратегии методы позволяют обучиться оптимальной стохастической стратегии

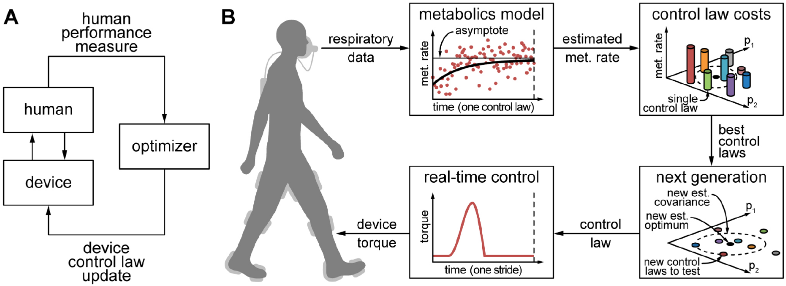
Функция полезности стратегии

- Цель: по данной стратегии $\hat{\pi}(s, a, \mathrm{w})$ с параметрами w, найти наилучшее значение w
- Как оценить качество стратегии $\hat{\pi}(\mathrm{w})$?
- В эпизодических средах мы можем использовать **начальную полезность**:
$$J_1(\mathrm{w}) = V^{\hat{\pi}(\mathrm{w})}(s_1) = \mathbb{E}_{\hat{\pi}(\mathrm{w})}[R_1]$$
- В непрерывных средах мы можем использовать **среднюю полезность**:
$$J_{avV}(\mathrm{w}) = \sum_s d^{\hat{\pi}(\mathrm{w})}(s) V^{\hat{\pi}(\mathrm{w})}(s)$$
- $d^{\hat{\pi}(\mathrm{w})}(s)$ – стационарное распределение марковской цепи для $\hat{\pi}(\mathrm{w})$
- Еще один вариант – среднее вознаграждение за шаг:
$$J_{avR}(\mathrm{w}) = \sum_s d^{\hat{\pi}(\mathrm{w})}(s) \sum_a \hat{\pi}(s, a, \mathrm{w}) \mathcal{R}_{sa}$$

Оптимизация стратегии

- Методы поиска стратегии – это алгоритмы оптимизации
- Необходимо найти значение w, которое максимизирует функционал $J(\mathrm{w})$ или $V^{\hat{\pi}(\mathrm{w})}$
- Есть ряд методов, не использующих градиентный спуск:
    - восхождение по выпуклой поверхности (hill climbing),
    - симплекс метод (simplex) или метод Нелдера-Мида,
    - генетические алгоритмы
- Иногда могут демонстрировать отличную производительность (например, эволюционные алгоритмы как альтернатива классическому RL)

Неградиентеные методы: экзоскелет



Оптимизация выполнялась с помощью метода CMA-ES - варианта адаптации матрицы ковариаций (Zhang et al. Science 2017)

Оптимизация стратегии

- Однако большей эффективности можно добиться с помощью градиентных методов:
  - градиентный спуск,
  - метод сопряженный градиентов (conjugate gradient),
  - квази-ньютоновские методы (quasi-newton)
- Для нас наибольший интерес представляет градиентный спуск с большим количеством вариаций
- Будем иметь в виду эпизодический марковский процесс принятия решений

Траектория

- Посчитаем градиент стратегии **аналитически**
- Предположим, что $\hat{\pi}(\mathrm{w})$ дифференцируема, если она отлична от 0
- И мы можем вычислить градиент $\nabla_{\mathrm{w}}\hat{\pi}(s, a, \mathrm{w})$
- Будем называть траекторией следующую цепочку состояний-действий:

$$\tau = (s_0, a_0, r_0, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$$

- Пусть $r(\tau) = \sum_{t=0}^{T} r(s_t, a_t)$ - сумма вознаграждений по траектории τ

Стратегия и отношение правдоподобия

- Полезность стратегии будет определяться как

$$J(\mathrm{w}) = \mathbb{E}_{\pi(\mathrm{w})}\left[\sum_{t=0}^{T} r(s_t, a_t)\right] = \sum_{\tau} p(\tau, \mathrm{w}) r(\tau).$$

- Здесь $p(\tau, \mathrm{w})$ - распределение вероятностей по траекториям τ при стратегии $\pi(\mathrm{w})$
- Оптимизационная задача запишется следующим образом:

$$\arg\max_{\mathrm{w}} J(\mathrm{w}) = \arg\max_{\mathrm{w}} \sum_{\tau} p(\tau, \mathrm{w}) r(\tau)$$

- Наша задача - найти параметры w стратегии, распишем градиент:

$$\nabla_{\mathrm{w}} J(\mathrm{w}) = \nabla_{\mathrm{w}} \sum_{\tau} p(\tau, \mathrm{w}) r(\tau) =$$

$$= \sum_{\tau} \nabla_{\mathrm{w}} p(\tau, \mathrm{w}) r(\tau) = \sum_{\tau} \frac{p(\tau, \mathrm{w})}{p(\tau, \mathrm{w})} \nabla_{\mathrm{w}} p(\tau, \mathrm{w}) r(\tau) =$$

$$= \sum_{\tau} p(\tau, \mathrm{w}) r(\tau) \frac{\nabla_{\mathrm{w}} p(\tau, \mathrm{w})}{p(\tau, \mathrm{w})} = \sum_{\tau} p(\tau, \mathrm{w}) r(\tau) \nabla_{\mathrm{w}} \log p(\tau, \mathrm{w})$$

Стратегия и отношение правдоподобия

- Оптимизационная задача запишется следующим образом:

$$\arg\max_{\mathrm{w}} J(\mathrm{w}) = \arg\max_{\mathrm{w}} \sum_{\tau} p(\tau, \mathrm{w}) r(\tau)$$

- Градиент по w:

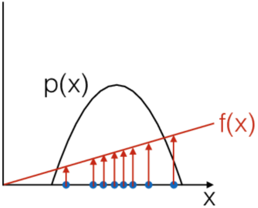
$$\nabla_{\mathrm{w}} J(\mathrm{w}) = \sum_{\tau} p(\tau, \mathrm{w}) r(\tau) \nabla_{\mathrm{w}} \log p(\tau, \mathrm{w})$$

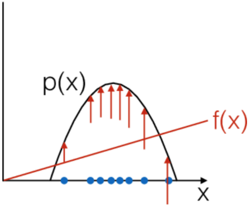
- В качестве приближения - эмпирическая оценка по выборке размера m:

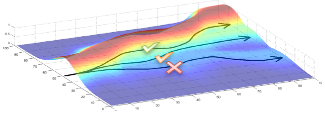
$$\nabla_{\mathrm{w}} J(\mathrm{w}) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} r(\tau^{(i)}) \nabla_{\mathrm{w}} \log p(\tau^{(i)}, \mathrm{w})$$

Результирующая функция

- Общий вид для $r(\tau^{(i)})\nabla_{\mathrm{w}} \log p(\tau^{(i)}, \mathrm{w})$: $\hat{g}_i = f(x_i)\nabla_{\mathrm{w}} \log p(x_i, \mathrm{w})$
- $f(x)$ измеряет насколько полезен пример x
- Сдвигаясь в направлении $\hat{g}_i$, увеличиваем $\log p$ примера пропорционально его полезности
- Это справедливо и для неизвестной функции $f(x)$ и дискретного множества примеров







Результирующая функция

- Эмпирическая оценка полезности по выборке размера m

$$\nabla_{\mathrm{w}} J(\mathrm{w}) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} r(\tau) \nabla_{\mathrm{w}} \log p(\tau^{(i)}, \mathrm{w})$$

- Пусть $\mu(s_0)$ - распределение начальных состояний, тогда

$$\nabla_{\mathrm{w}} \log p(\tau^{(i)}, \mathrm{w}) = \nabla_{\mathrm{w}} \log \left[\mu(s_0) \prod_{t=0}^{T-1} \hat{\pi}(a_t, s_t, \mathrm{w}) P(s_{t+1} | s_t, a_t) \right] =$$
$$= \nabla_{\mathrm{w}} \left[\log \mu(s_0) + \sum_{t=0}^{T-1} \log \hat{\pi}(a_t, s_t, \mathrm{w}) + \log P(s_{t+1} | s_t, a_t) \right] =$$
$$= \sum_{t=0}^{T-1} \log \hat{\pi}(a_t, s_t, \mathrm{w})$$

- Результирующая функция (score function) – это $\nabla_{\mathrm{w}} \log \hat{\pi}_{\mathrm{w}}(s, a, \mathrm{w})$

Градиент стратегии для эпизодической среды

- Оптимизационная задача запишется следующим образом:

$$\arg\max_{\mathrm{w}} J(\mathrm{w}) = \arg\max_{\mathrm{w}} \sum_{\tau} p(\tau, \mathrm{w}) r(\tau)$$

- Эмпирическая оценка полезности по выборке размера m для стратегии $\pi(\mathrm{w})$ по результирующей функции:

$$\nabla_{\mathrm{w}} J(\mathrm{w}) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} r(\tau) \nabla_{\mathrm{w}} \log p(\tau^{(i)}, \mathrm{w}) =$$

$$= \frac{1}{m} \sum_{i=1}^{m} r(\tau^{(i)}) \sum_{t=0}^{T-1} \nabla_{\mathrm{w}} \log \pi(a_t^{(i)}, s_t^{(i)}, \mathrm{w})$$

- Нам не нужна модель динамики!
- Несмещенная, но очень зашумленная оценка

Теорема о градиенте стратегии

- Теорема о градиенте стратегии обобщает подход коэффициентов правдоподобия.
- Заменяем текущее значение отдачи на долговременное значение оценки полезности $Q^{\hat{\pi}(\mathrm{w})}(s, a)$
- Теорема о градиенте стратегии применяется к полной отдаче, среднему вознаграждению и средней полезности

Theorem

Для любой дифференцируемой стратегии $\hat{\pi}(s, a, \mathrm{w})$, для любой функции полезности стратегии $J = J_1$, J_{avR} или $\frac{1}{1-\gamma} J_{avV}$ градиент стратегии равен:

$$\nabla_{\mathrm{w}} J(\mathrm{w}) = \mathbb{E}_{\hat{\pi}(\mathrm{w})}[\nabla_{\mathrm{w}} \log \hat{\pi}(s, a, \mathrm{w}) Q^{\hat{\pi}(\mathrm{w})}(s, a)]$$

Монте-Карло градиент стратегии

- Будем использовать отдачу R_t в качестве несмещенной оценки $Q^{\hat{\pi}(\mathrm{w})}(s, a)$:

$$\nabla_{\mathrm{w}} \mathbb{E}[r(\tau)] \approx \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{T-1} \nabla_{\mathrm{w}} \log \hat{\pi}(s, a, \mathrm{w}) R_t$$

- Обновляем параметры с помощью стохастического градиентного спуска.

Algorithm 1 REINFORCE

1: **function** REINFORCE
2: инициализируем w;
3: **for** каждого эпизода $\{s_1, a_1, r_2, \dots, s_{T-1}, a_{T-1}, r_T\} \sim \hat{\pi}(\mathrm{w})$ **do**
4: **for** $t = 1$ и до $T - 1$ **do**
5: $\mathrm{w} \leftarrow \mathrm{w} + \alpha \nabla_{\mathrm{w}} \log \hat{\pi}(s_t, a_t, \mathrm{w}) R_t$
return w

Пример параметризации: логистическая стратегия

- Будем использовать логистическую стратегию в качестве примера.
- Взвесим действия, используя линейную комбинацию признаков $\phi(s,a)^T\mathrm{w}$.
- Вероятность действия пропорциональна экспоненциальным весам:

$$\hat{\pi}(s,a,\mathrm{w}) = \frac{e^{\phi(s,a)^T\mathrm{w}}}{\sum_a e^{\phi(s,a)^T\mathrm{w}}}.$$

- Результирующая функция:

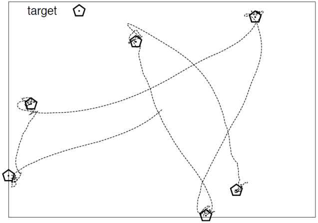
$$\nabla_{\mathrm{w}} \log \hat{\pi}(s,a,\mathrm{w}) = \phi(s,a) - \mathbb{E}_{\hat{\pi}(\mathrm{w})}[\phi(s,\cdot)].$$

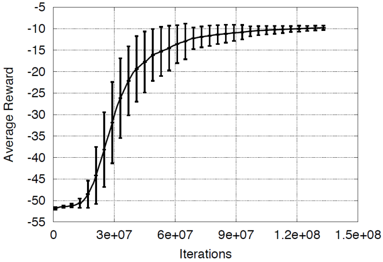
Гауссова стратегия

- В непрерывном пространстве действий подойдет гауссова стратегия.
- Среднее – это линейная комбинация признаков состояния $\mu(s) = \phi(s)^T \mathrm{w}$.
- Дисперсия может быть фиксированной σ^2, или тоже может быть параметризована.
- Гауссова стратегия задается как $\approx \mathcal{N}(\mu(s), \sigma^2)$.
- Результирующая функция:

$$\nabla_{\mathrm{w}} \log \hat{\pi}(s, a, \mathrm{w}) = \frac{(a - \mu(s))\phi(s)}{\sigma^2}.$$

Пример с шайбой





- Непрерывное пространство действий – оказание небольшого воздействия на шайбу.
- Мы получаем вознаграждение, когда шайба оказалась возле цели.
- Положение цели меняется каждые 30 секунд.
- Стратегия была построена с использованием одного из вариантов Монте-Карло градента стратегии.

Требования к градиенту стратегии

- Цель - максимально быстро сойтись к локальному минимуму
  - Вычисляя вознаграждения при выполнении стратегии, хотим минимизировать количество итераций до достижения нужной стратегии
- Во время поиска стратегии поочередно оцениваем стратегию и обновляем ее по аналогии с итерациями по стратегиям
- Основная задача - добиться максимального монотонного улучшения стратегии на каждой итерации:
  - получение более точной оценки градиента (улучшение обновления параметров стратегии),
  - изменение способа обновления параметров стратегии на основе полученного градиента

Оценка градиента стратегии

- Оценка градиента по траекториям:

$$\nabla_\theta J(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} r(\tau^{(i)}) \sum_{i=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)}).$$

- Оценка несмещенная, но с высокой дисперсией
- Способы борьбы с дисперсией:
  - использование временной структуры MDP,
  - **учет базового уровня**,
  - использование других оценок вместо Монте-карло подхода

Базовый уровень для градиента стратегии

- Уменьшение дисперсии введением базового уровня $B(s_t)$:

$$\nabla_\theta \mathbb{E}_\tau[R(\tau)] = \mathbb{E}_\tau \left[\sum_{i=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)}) \left(\sum_{t'=t}^{T-1} r_{t'} - B(s_t) \right) \right]$$

- При любом выборе $B(S_t)$ оценка градиента останется несмещенной
- Квази-оптимальный выбор базового уровня - ожидаемая отдача:

$$B(s_t) \approx [r_t + r_{t+1} + \ldots r_{T-1}]$$

- Интерпретация: увеличение $\log p$ действия a_t пропорционально тому, насколько отдача $\sum r_{t'}$ лучше ожидаемой

Базовый алгоритм градиента стратегии

Algorithm 2 Vanilla PG

function VPG
    Инициализируем параметры стратегии θ и базовый уровень B
    for итераций $i = 1, 2, \dots,$ **do**
        набираем множество траекторий, выполняя текущую стратегию
        for каждого шага t траектории τ^i **do**
            $R_t^i = \sum_{t'=t}^{T-1} r_{t'},$
            $\hat{A}_t^i = R_t^i - B(s_t)$ - **оценка преимущества**,
        обновляем базовый уровень, минимизируя $\sum_i \sum_t \|B(s_t) - R_t^i\|^2,$
        $\hat{g} = \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}_t,$
        обновляем стратегию, используя оценку градиента $\hat{g}$.

Выбор базового уровня

- Функция полезности состояния-действия:

$$Q^{\pi,\gamma}(s,a) = \mathbb{E}_\pi[r_0 + \gamma r_1 + \gamma^2 r_2 + \dots | s_0 = s, a_0 = a]$$

- Функция полезности состояния может служить отличным базовым уровнем:

$$V^{\pi,\gamma}(s) = \mathbb{E}_\pi[r_0 + \gamma r_1 + \gamma^2 r_2 + \dots | s_0 = s] = \mathbb{E}_{a\sim\pi}[Q^{\pi,\gamma}(s,a)]$$

- **Функция преимущества** (advantage function) - комбинация функций полезности и базового уровня:

$$A^{\pi,\gamma} = Q^{\pi,\gamma}(s,a) - V^{\pi,\gamma}(s)$$

Оценка градиента стратегии

- Оценка градиента по траекториям:

$$\nabla_\theta J(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)}) \sum_{i=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}).$$

- Оценка несмещенная, но с высокой дисперсией.
- Способы борьбы с дисперсией:
  - использование временной структуры MDP,
  - учет базового уровня,
  - **использование других оценок вместо Монте-карло подхода**.

Уменьшение дисперсии с использованием критика

- В методе Монте-Карло градиента стратегии большая дисперсия решения
- Попробуем использовать **критика** (critic) для оценки функции полезности действия:

$$Q_w(s, a) \approx Q^{\pi_\theta}(s, a)$$

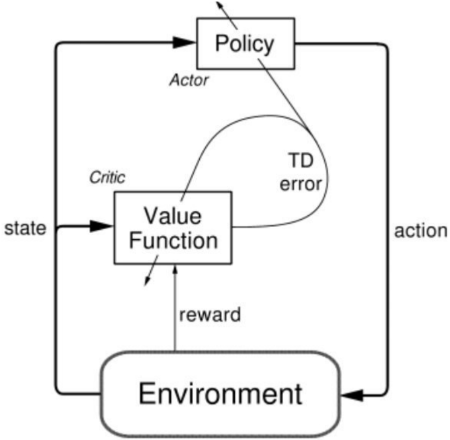
- Алгоритмы актор-критик (actor-critic) поддерживают два множества параметров:
  - критик — обновляет параметры w функции полезности действия,
  - актор — обновляет параметры θ стратегии с учетом предположений критика
- Алгоритмы актор-критик следуют по градиенту **приближенной** стратегии:

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)],$$

$$\Delta\theta = \alpha \nabla_\theta \log \pi_\theta(s, a) Q_w(s, a).$$

Оценка функции полезности действия

- Критик оценивает стратегию
- Насколько хороша стратегия π_θ при текущих параметрах θ
- Мы знаем следующие способы решения этой задачи:
  - Монте-Карло оценка стратегии,
  - обучение на основе временных различий
- Можем также использовать оценку стратегии методом наименьших квадратов



Полезность действия актор-критика

- Рассмотрим простейший алгоритм актор-критика на основе полезности действия критика
- Будем использовать линейную аппроксимационную функцию $Q_w(s,a) = \phi(s,a)^T w$:

  критик обновляет параметры w с помощью TD(0),

  актор обновляет параметры θ, используя градиент стратегии

Algorithm 3 QAC

1: **function** QAC
2: Инициализируем s, θ
3: Выбираем $a \sim \pi_\theta$
4: **for all** шагов **do**
5: получаем вознаграждение $r = \mathcal{R}_s^a$ и следующее состояние $s' \sim \mathcal{P}_s^a$
6: выбираем следующее действие $a' \sim \pi_\theta(s')$
7: $\delta = r + \gamma Q_w(s', a') - Q_w(s, a)$
8: $\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)$
9: $w \leftarrow w + \beta \delta \phi(s, a)$
10: $a \leftarrow a', s \leftarrow s'$

Смещенность в алгоритмах актор-критик

- Аппроксимация градиента стратегии приводит к смещению оценки
- Смещенный градиент стратегии может не позволить найти правильное решение
- Например, использование признаков в $Q_w(s, a)$ для клеточного мира с неоднозначным определением состояния
- К счастью, у нас есть возможность выбрать такую функцию аппроксимации, которая позволит избежать смещенных оценок
- Можем все-еще использовать точный градиент стратеги

Совместимые функции аппроксимации

Theorem (о совместимой функции аппроксимации)

Если удовлетворены следующие два условия:

1. *аппроксиматор функции полезности* ***совместим*** *со стратегией:*

$$\nabla_w Q_w(s, a) = \nabla_\theta \log \pi_\theta(s, a),$$

2. *параметры функции полезности минимизируют среднеквадратичную ошибку:*

$$\epsilon = \mathbb{E}_{\pi_\theta}[(Q^{\pi_\theta}(s, a) - Q_w(s, a))^2],$$

тогда градиент стратегии в точности равен

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)].$$

Оценка функции преимущества

- Функция преимущества (advantage function) может существенно уменьшить дисперсию градиента стратегии
- Критик должен на самом деле оценивать функцию преимущества
- Например, оценивая как $V^{\pi_\theta}(s)$, так и $Q^{\pi_\theta}(s, a)$
- Используем два аппроксиматора и два вектора параметров:

$$V_v(s) \approx V^{\pi_\theta}(s),$$

$$Q_w(s, a) \approx Q^{\pi_\theta}(s, a),$$

$$A(s, a) = Q_w(s, a) - V_v(s)$$

- Обновляем обе функции полезности с помощью, например, TD-обучения

Оценка функции преимущества

- Для истинной функции полезности $V^{\pi_\theta}(s)$, TD-ошибка равна

$$\delta^{\pi_\theta} = r + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s).$$

- Она является несмещенной оценкой функции преимущества:

$$\begin{aligned}
\mathbb{E}_{\pi_\theta}[\delta^{\pi_\theta}|s,a] &= \mathbb{E}_{\pi_\theta}[r + \gamma V^{\pi_\theta}(s')|s,a] - V^{\pi_\theta}(s)\\
&= Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\\
&= A^{\pi_\theta}(s,a).
\end{aligned}$$

- Таким образом, мы можем использовать TD-ошибку для вычисления градиента стратегии:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\delta^{\pi_\theta}].$$

- На практике, мы используем аппроксимацию TD-ошибки:

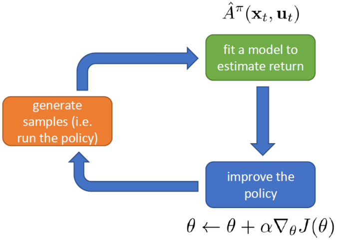
$$\delta_v = r + \gamma V_v(s') - V_v(s).$$

- В этом подходе нам достаточно использовать только один вектор параметров v

Общая схема алгоритмов градиента стратегии

Algorithm 4 Common template PG

1: **for** итераций $1, 2, 3, \ldots$ **do**
2: запустить стратегию для генерации T временных шагов или N траекторий,
3: на каждом шаге каждой траектории вычислить полезность $Q^{\pi}(s_t, a_t)$ и базовый уровень $B(s_t)$,
4: вычислить оценку градиента стратегии $\hat{g}$,
5: обновить стратегию на основе, возможно, с ограничением на локальную область.



Семейство алгоритмов градиента стратегии

- Градиент стратегии имеет несколько эквивалентных формы:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) R_t] && \textit{REINFORCE} \\
&= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) Q^w(s,a)] && \textit{Q Actor} - \textit{Critic} \\
&= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) A^w(s,a)] && \textit{Advantage Actor} - \textit{Critic} \\
&= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) \delta] && \textit{TD Actor} - \textit{Critic} \\
G_\theta^{-1} \nabla_\theta J(\theta) &= w && \textit{Natural Actor} - \textit{Critic}
\end{aligned}
$$

- Каждый из может быть использован в стохастическом градиентном спуске.
- Критик использует оценку стратегии (MC или TD-обучение) для оценки $Q^\pi(s,a), A^\pi(s,a), V^\pi(s)$.

Выбор оценки полезности траектории

- R_t^i - оценка функции полезности состояния на основе одного прогона (roll out).
- Эта оценка несмещенная, но обладает высокой дисперсией.
- Уменьшение дисперсии за счет введения смещения (временные различия и аппроксимация функции).
- Оценка V/Q делается **критиком**.
- Методы **актор-критика** поддерживают явное представление и стратегии, и функции полезности.
- Пример - метод A3C (Asynchronous Advantage Actor-Critic) `https://arxiv.org/abs/1602.01783` - один из самых популярных в настоящее время, поддерживает параллельные вычисления

Градиент стратегии с функцией полезности

- Оценка полезности траектории

$$\nabla_\theta \mathbb{E}_\tau[R(\tau)] = \mathbb{E}_\tau \left[\sum_{i=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \left(\sum_{t'=t}^{T-1} r_{t'} - B(s_t) \right) \right]$$

$$\nabla_\theta \mathbb{E}_\tau[R(\tau)] = \mathbb{E}_\tau \left[\sum_{i=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \left(Q_w(s_t) - B(s_t) \right) \right]$$

- Если наш базовый уровень определяется функцией полезности, мы получаем использование функции преимущества:

$$\nabla_\theta \mathbb{E}_\tau[R(\tau)] = \mathbb{E}_\tau \left[\sum_{i=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \hat{A}^\pi(s_t, a_t) \right]$$

Оценка полезности траектории: N-шаговые оценки

$$\nabla_\theta V(\theta) \approx (1/m) \sum_{i=1}^{m} \sum_{i=0}^{T-1} R_t^i \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)})$$

- Критик может выбрать любую смесь между оценками TD и MC для целевого значения, чтобы заменить истинную функцию полезности состояния-действия:

$$\hat{R}_t^{(1)} = r_t + \gamma V(s_{t+1})$$
$$\hat{R}_t^{(2)} = r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2})$$
$$\hat{R}_t^{(\text{inf})} = r_t + \gamma r_{t+1} \gamma^2 r_{t+2} + \dots$$

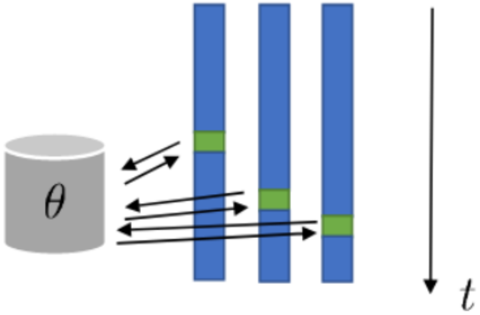
- По аналогии получаем с функцией преимущества:

$$\hat{A}_t^{(1)} = r_t + \gamma V(s_{t+1}) - V(s_t)$$
$$\hat{A}_t^{(\text{inf})} = r_t + \gamma r_{t+1} \gamma^2 r_{t+2} + \dots - V(s_t)$$

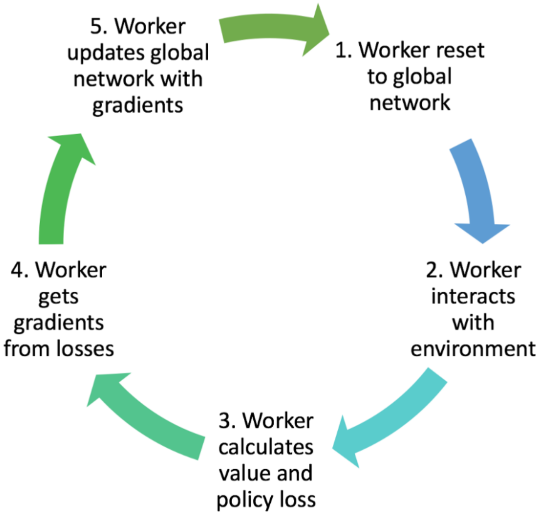
- $A_t^{(1)}$ имеет меньшую дисперсию и большую смещенность, $A_t^{(\text{inf})}$ - наоборот.

Особенности A3C

- Критик обновляет функцию полезности пока множество акторов работают параллельно
- Критики синхронизируются время от времени по глобальным переменным
- Использование *n*-шаговых оценок полезности



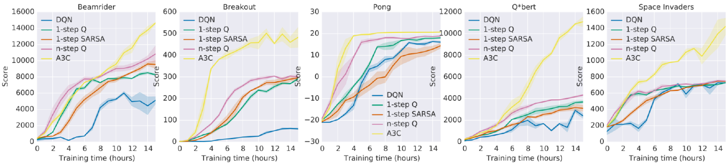
Цикл потоков в A3C



Алгоритм A3C

Algorithm 5 A3C

θ, w – глобальные переменные, θ', w' – для каждого потока, $t = 1$
while $T \leq T_{max}$ **do**
  $\Delta\theta = 0, \Delta w = 0$
  синхронизируем $\theta' = \theta, w' = w$
  $t_{start} = t$, выбираем состояние s_t
  while s_t - нетерминальное, $t - t_{start} < t_{max}$ **do**
    применяем $a_t \sim \pi_{\theta'}(a_t|s_t)$ и получаем r_t, s_{t+1}
    $t \leftarrow t + 1, T \leftarrow T + 1$
  $R = 0$ или $R = V_{w'}(s_t)$, если s_t - нетерминальное
  for $i = \{t - 1, \ldots, t_{start}\}$ **do**
    $r(\tau) \leftarrow \gamma r(\tau) + r_i$
    $\Delta\theta \leftarrow \theta + \nabla_{\theta'} \log \pi_{\theta'}(a_i|s_i)(r(\tau) - V_{w'}(s_i))$
    $\Delta w \leftarrow \Delta w + \nabla_{w'} (r(\tau) - V_{w'}(s_i))^2$



Особенности сочетания отложенных прецедентов и градиента стратегии

- Одновременно обновляем и функцию полезности Q и стратегию π
- Используем обучение по отложенном опыте и временные различия для полезности Q
- Тесно связна с Q-обучением, но работает для непрерывного пространства действий, решая оптимизационную задачу $\max_a \hat{Q}(s, a, w) \approx Q(s, \pi(s))$
- Необходимо поддерживать исследование среды, добавлением случайности в детерминированное действие - гауссов шум
- Алгоритм глубокий градиент детерминированной стратегии (DDPG)– глубокая Q-сеть для непрерывного пространства действий

Алгоритм DDPG

- Пусть D – наша память прецедентов (s, a, r, s'), она должна быть достаточно большой, чтобы избежать переобучения
- Функция потерь критика стандартная квадратичная:

$$L(w, D) = \mathbb{E}_{(s,a,r)\sim D}\left[\left(Q(s, a, w) - (r + \gamma \max_{a'} Q(s', a', w)\right)^2\right]$$

- Используем «замораживания» весов для показателя $r + \gamma \max_{a'} Q(s', a', w^-)$:

$$w^- \leftarrow \rho w^- + (1 - \rho) w$$

- Аналогично используем «замораживание» для стратегии – целевая стратегию $\mu(\theta^-)$, которая только примерно максимизирует $Q(w^-)$, для выбора $a' \sim \mu(\theta^-)$
- Находим детерминированную стратегию $\mu(s, \theta)$ градиентным спуском по дифференцируемой по a функции полезности:

$$\max_\theta \mathbb{E}_{s\sim D}\left[Q(s, \mu(s, \theta), w)\right]$$

Нормализация функции преимущества

- Ограничим вид функции $Q(s, a) = V(s) + A(s, a)$, чтобы задача оптимизации была тривиальной.
- Для примера - парабола для одномерного пространства действий:

$$A(s, a) = -k_\theta(s)(a - \mu_\theta(s))^2.$$

- Оптимальное действие - $a^* = \mu_\theta(s)$.
- В многомерном случае:

$$A(s, a) = -\frac{1}{2}(a - \mu_\theta(s))L(s)L(s)^\intercal(a - \mu_\theta(s))$$

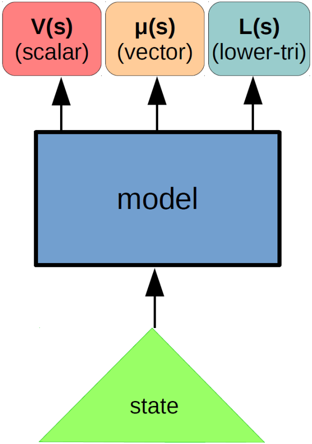
где $L(s)$ - нижнетреугольная матрица.

Аппроксимация при сквозном обучении

$$Q(s,a) = V(s) + A(s,a)$$

$$A(s,a) = -\frac{1}{2}(a - \mu_\theta(s))L(s)L(s)^\intercal(a - \mu_\theta(s))$$

$$\arg\min_\theta (Q(s_t,a_t) - [r + \gamma V(s_{t+1})])^2$$



Градиент детерминированной стратегии

- Будем обучать отдельный аппроксиматор для поиска a^*.
- Обучение критика $Q(s,a)$:

$$\arg\min_\theta (Q(s,a) - [r + \gamma Q(s_{t+1}, \mu_\theta(s_{t+1}))])^2.$$

- Обучение актора $a^*(s) \approx \mu_\theta(s)$ - сдвигаем параметры стратегию в направлении изменения полезности состояние-действий:

$$\nabla_\theta J = \nabla_\theta \mu(s|\theta) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}.$$