Исследование пространства признаков в задаче обучения с подкреплением

студент: **Гринчук Алексей Валерьевич**[1]
научный руководитель: проф. **Оселедец Иван Валерьевич**[2]
в сотрудничестве с: prof. **Ronald Edward Parr**[3]

[1]Кафедра "Интеллектуальные системы"
Московский физико-технический институт

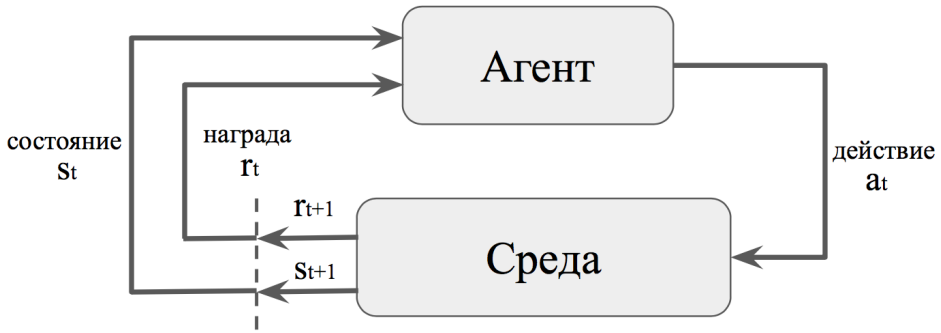
[2]Институт вычислительной математики РАН

[3]Computer Science Department
Duke University

15 июня 2017 г.

Марковский решающий процесс

- $s_t \in \mathcal{S}$ – состояние среды
- $a_t \in \mathcal{A}$ – выбранное действие
- $r_t \in \mathcal{R}$ – полученная награда



Задача обучения с подкреплением

Целью агента в задаче обучения с подкреплением является нахождение стратегии, действуя согласно которой он может получить максимальную суммарную награду.

Более формально: найти отображение $\pi(a|s) = \mathbb{P}(a_t = a|s_t = s)$ из множества состояний в множество вероятностных распределений выбора действий (так называемая *стратегия* агента), которая максимизирует суммарную ожидаемую дисконтируемую награду.

$$\mathbb{E}_\pi(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots) \to \max_\pi$$

Изображение на слайде было взято из книги *Sutton & Barto, Reinforcement Learning: An Introduction (1998)*.

Q-функция и уравнение Беллмана

- Стратегия агента $\pi(a|s) = \mathbb{P}(a_t = a|s_t = s)$
- Матрица вероятностей переходов $P(s, a, s') = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a)$
- Q-функция $Q^\pi(s, a) = \mathbb{E}_\pi(r_t + \gamma r_{t+1} + \ldots|s_t = s, a_t = a)$
- Уравнение Беллмана:

$$Q^\pi(s, a) = R(s, a) + \gamma \sum_{s'} P(s, a, s') \sum_{a'} \pi(a'|s') Q^\pi(s', a')$$

- ВП для пар состояние-действие $P^\pi(s', a'|s, a) = P(s, a, s')\pi(a'|s')$

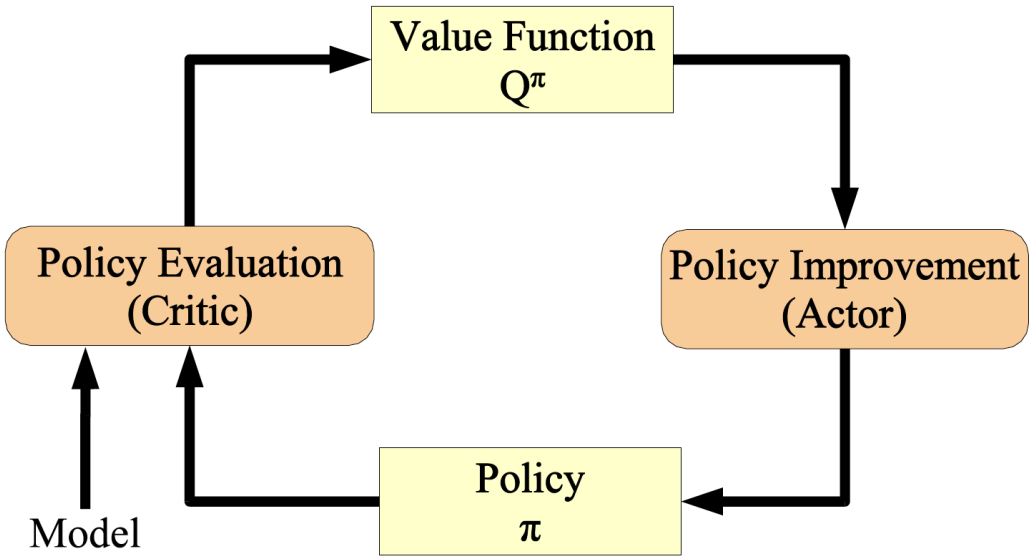
Неподвижная точка

- Оператор Беллмана:

$$(TQ^\pi)(s, a) = R(s, a) + \gamma \sum_{s', a'} P^\pi(s, a|s', a') Q^\pi(s', a')$$

- Q^π является неподвижной точкой оператора Беллмана: $TQ^\pi = Q^\pi$

Как решать задачи обучения с подкреплением?



- Улучшение стратегии — найти стратегию по Q-функции:

$$\pi_{i+1}(s) = \arg\max_{a \in \mathcal{A}} Q^{\pi_i}(s, a).$$

- Оценка стратегии — найти Q-функцию по стратегии:

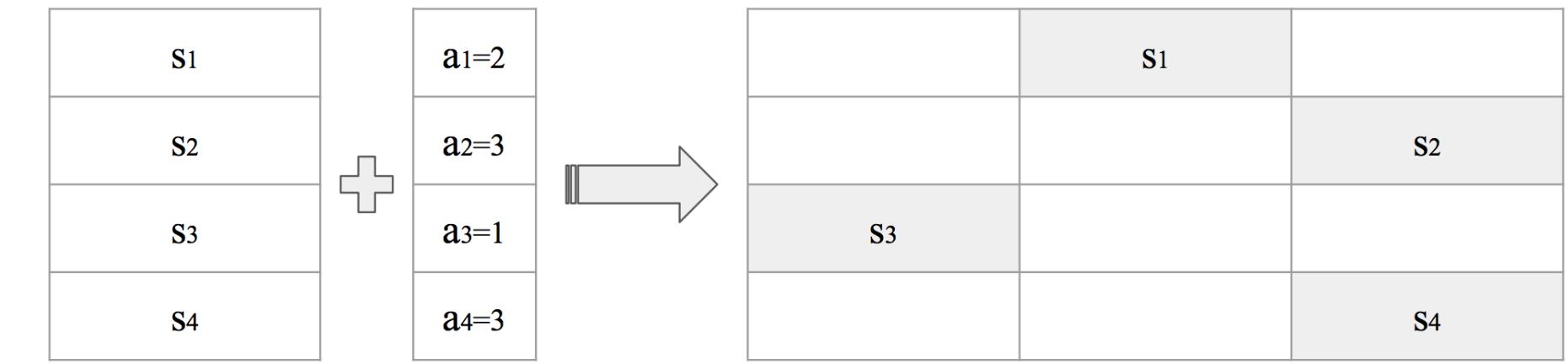
$$TQ^{\pi_i} = Q^{\pi_i}.$$

Изображение на слайде было взято из статьи *Lagoudakis & Parr. Least-Squares Policy Iteration (2003).*

Обучение по истории агента

На практике у нас обычно нет доступа к $P(s, a, s')$, однако у нас имеется доступ к истории взаимодействий агента со средой: $\{(s, a, r, s')\}$, сгенерированные исходя из стратегии $\pi(a|s)$ (например, случайной). В данной работе мы используем следующие матричные обозначения:

- $A \in \mathbb{R}^{n \times lm}$ – матрица пар состояние-действие (s, a)
- $P^{\pi} A = A' \in \mathbb{R}^{n \times lm}$ – матрица следующих пар состояние-действие (s', a')
- $R \in \mathbb{R}^{n}$ – вектор наград



состояния и соответствующие действия

матрица пар состояние-действие A

Аппроксимация Q-функции

В некоторых случаях Q-функция может быть получена в виде точного решения уравнения Беллмана, однако в общем случае это невозможно.

Линейная аппроксимация Q-функции

- Мы аппроксимируем Q-функцию линейной комбинацией признаков:

$$\hat{Q}^\pi = A\mathbf{w}_A^\pi,$$

  где $\mathbf{w}_A^\pi \in \mathbb{R}^{ml}$ — вектор весов.
- Линейные методы ищут решение $\mathbf{w}_A^\pi$ как решение следующего уравнения неподвижной точки:

$$A\mathbf{w}_A^\pi = \Pi(R + \gamma A'\mathbf{w}_A^\pi),$$

  где $\Pi = A(A^TA)^{-1}A^T$ — это ортогональный l_2 проектор на $\text{span}(A)$.
- Решая приведённое выше уравнение получаем:

$$\mathbf{w}_A^\pi = (A^TA - \gamma A^TA')^{-1}A^TR.$$

Линейное сжатие признаков

Проблема

Сложность подсчёта вектора весов есть $\mathcal{O}([lm]^3)$, что делает невозможным решение современных практических задач с $ml \approx 10^4 - 10^5$.

Идея решения

Состояния, представленные в виде картинок, имеют слишком много признаков (пикселей). Будем искать линейное преобразование (кодировщик) $E^\pi \in \mathbb{R}^{lm \times k}$ из исходного пространства высокой размерности в пространство низкой размерности.

Предсказание признаков следующих состояний

- $\Phi = AE^\pi \in \mathbb{R}^{n \times k}$ — матрица пар состояние-действие в пространстве низкой размерности
- $\hat{Q}^\pi = \Phi \mathbf{w}_\Phi^\pi$ — аппроксимация Q-функции
- $\mathbf{w}_\Phi^\pi = (\Phi^T \Phi - \gamma \Phi^T \Phi')^{-1} \Phi^T R$ — вектор весов

Теорема

Если существуют две матрицы E^π и D^π такие что:

$$AE^\pi D^\pi = [R, A' E^\pi],$$

то существует вектор весов $\mathbf{w}$ такой что $\hat{Q}^\pi = AE^\pi \mathbf{w} = \Phi \mathbf{w}$ является неподвижной точкой оператора Беллмана $TQ^\pi = Q^\pi$.

На практике мы минимизируем норму Фробениуса разности:

$$\|AE^\pi D^\pi - [R, A' E^\pi]\|_F \to \min_{E^\pi, D^\pi} \tag{1}$$

Compressed Value Iteration

1. Минимизировать (1) для $k = 1$ (научиться предсказывать вектор наград):

$$AE_0D_0 = R \Rightarrow E_0D_0 = A^\dagger R \Rightarrow E_0 = A^\dagger R,\ D_0 = 1.$$

2. Минимизировать (1) для произвольного k, если оно было минимизировано для всех $p < k$ (научиться предсказывать вектор наград и признаки следующих состояний):

$$AE_kD_k = [R, A'E_{k-1}] \Rightarrow X_k = E_kD_k = A^\dagger[R, A'E_{k-1}],$$

$$E_k, D_k = \mathrm{QR}(X_k).$$

3. Найти вектор весов $\mathbf{w}$, аппроксимацию Q-функции $\hat{Q} = AE_k\mathbf{w}$, и стратегию $\pi(a|s) = \begin{cases} 1, & a = \arg\max_{a \in \mathcal{A}} \hat{Q}(s,a), \\ 0, & \text{иначе.} \end{cases}$.

Подпространство Крылова

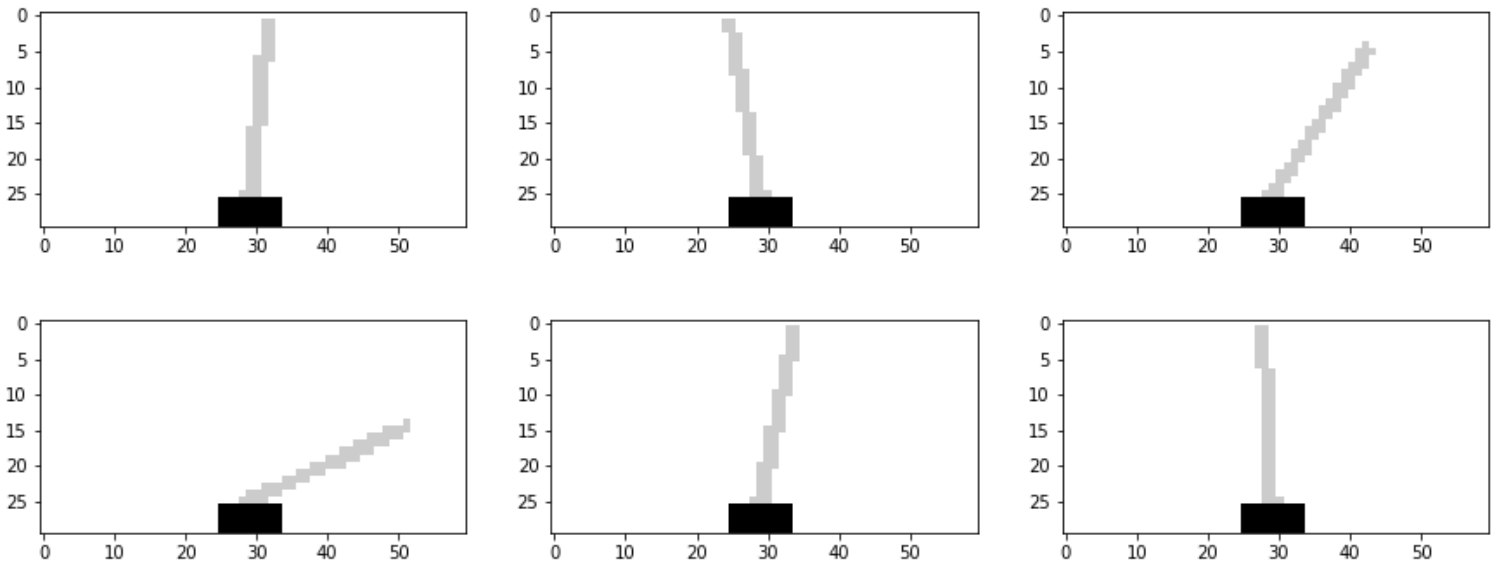
Лемма

Для любой матрицы $E^\pi \in \mathbb{R}^{ml \times k}$ и невырожденной матрицы Y, линейные аппроксимации Q-функции для двух различных наборов признаков $\Phi = AE^\pi$ и $\Phi_Y = \Phi Y = AE^\pi Y$ в точности совпадают: $\hat{Q}^\pi_\Phi \equiv \hat{Q}^\pi_{\Phi_Y}$.

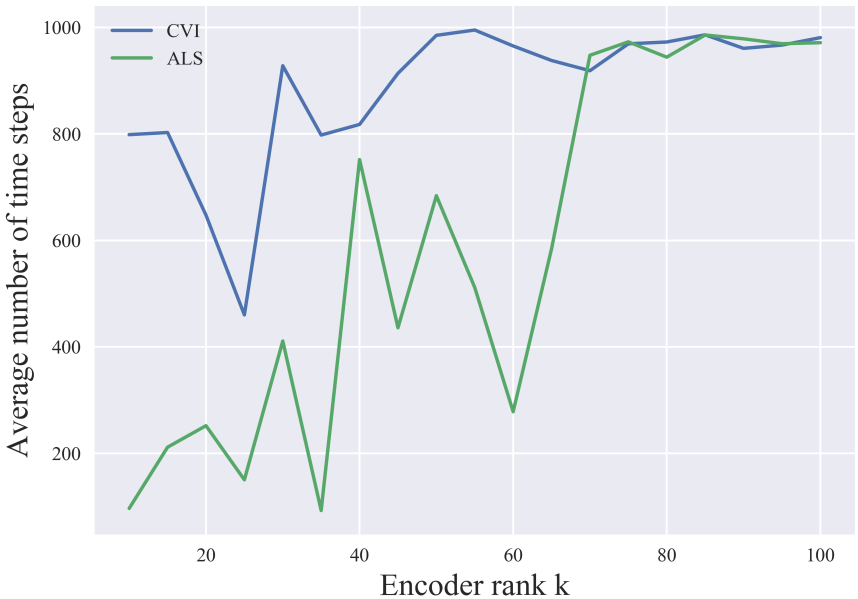
- $E^\pi_1 = A^\dagger R$
- $E^\pi_2 = A^\dagger [R, A' E^\pi_1] = [A^\dagger R,\ A^\dagger A' A^\dagger R]$
- $\ldots$
- $E^\pi_k = [A^\dagger R,\ A^\dagger A' A^\dagger R,\ \ldots,\ (A^\dagger A')^{k-1} A^\dagger R]$

Эксперимент: Перевёрнутый Маятник

- Каждый кадр является картинкой разрешения 30×60 пикселей, каждое состояние состоит из двух последовательных кадров, вытянутых в строку (вектор размерности 3600), каждая пара состояние-действие является вектором размерности 10800.
- Среднее количество временных шагов прежде чем маятник падает для случайной стратегии равно ≈ 11, максимальное количество равно ≈ 32.
- Во время тестирования стратегии, если маятник не упал после 1000 временных шагов, мы провозглашаем это успехом и прерываем тестовую симуляцию.

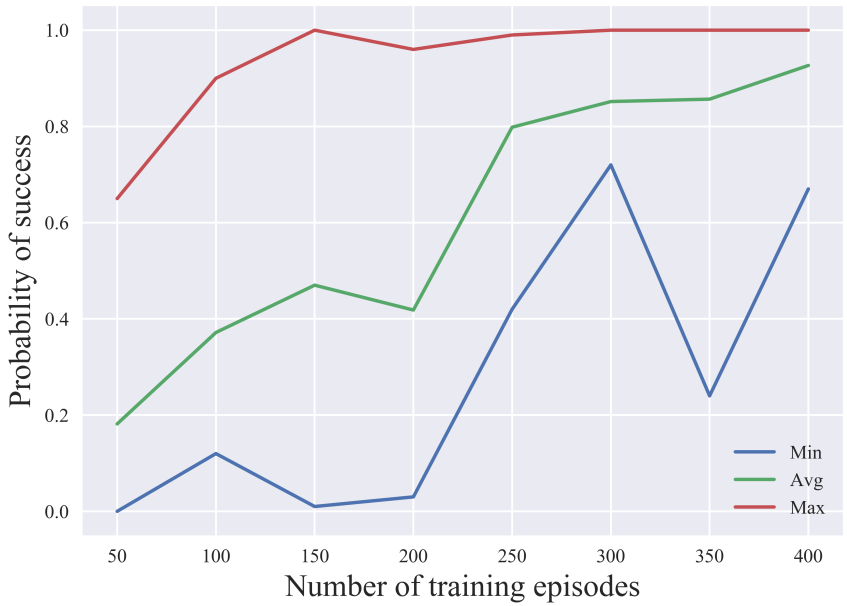


Эксперимент: Перевёрнутый Маятник



Количество балансирующих временных шагов и вероятность успеха в зависимости от количества признаков.

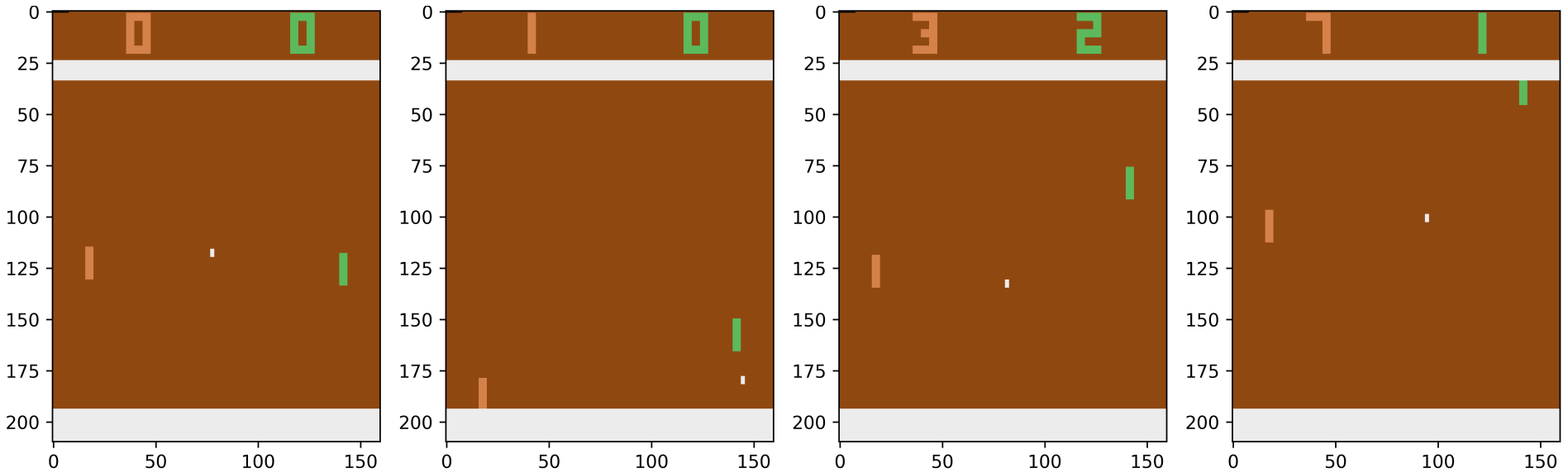
Эксперимент: Перевёрнутый Маятник



Количество балансирующих временных шагов и вероятность успеха в зависимости от размера обучающей выборки для CVI усреднённое для 6 различных обучающих выборок.

Эксперимент: Атари 2600 Пинг-Понг

- Каждый кадр является картинкой разрешения 80×80 пикселей, каждое состояние состоит из четырёх последовательных кадров, вытянутых в строку (вектор размерности 25600), каждая пара состояние-действие является вектором размерности 76800.
- Для каждого временного шага симуляции предусмотрена нагада 0. Наш агент получает награду 1, если он выигрывает очко и -1, если он его проигрывает.
- Мы формируем обучающую выборку посредством запуска симуляции для нужного количества временных шагов, выбирая действия случайно.



Эксперимент: Атари 2600 Пинг-Понг

Разреженность состояний

- Большая часть всех кадров — фон.
- После бинаризации кадров (ракетки и мячик — единички, фон — нолики) и добавления действий, мы получаем матрицы состояний-действий A and A', разреженные на 99.5%.
- Пакет Питона *scipy.sparse* позволяет работать с разреженными матрицами, что даёт 30-кратный выигрыш в скорости и 100-кратный выигрыш в памяти.

игра до 1 очка		игра до 2 очков		игра до 3 очков	
Random	CVI, $k = 50$	Random	CVI, $k = 35$	Random	CVI, $k = 50$
$< 3.5\%$	46%	$< 2.5\%$	37%	$< 1.5\%$	11%

Сравнение стратегий, полученных в результате работы CVI и случайной стратегии в смысле процента выигранных очков.

Заключение

- Доказано, что достаточно уметь предсказывать признаковые описания следующих состояний в пространстве низкой размерности и вектор наград, чтобы гарантировать оптимальность линейной аппроксимации Q-функции.
- Предложен *compressed value iteration* — алгоритм, который итеративно наращивает пространство признаков и решает оптимизационную задачу, порождённую идеей предсказания признаков следующих состояний.
- Предложена модификация алгоритма CVI, которая использует методы Крылова для генерации признаков. Также предложена другая модификация, завязанная на специфике состояний некоторых задач обучения с подкреплением и позволяющая получить выигрыш в скорости и памяти, если состояния сильно разрежены.
- Проведена серия вычислительных экспериментов на двух популярных задачах обучения с подкреплением “Перевёрнутый Маятник” и Атари 2600 “Пинг-Понг”, результаты которой подтверждают тот факт, что предложенный метод работает и эффективен.