RL: CrossEntropy

Обучение с учителем

- Задано:
  - Предиктор и таргет (x,y)
  - Семейство алгоритмов $a_\theta(x) \to y$
  - Лосс функция $L(y, a_\theta(x))$

- Найти:

$$\theta' \leftarrow argmin_\theta L(y, a_\theta(x))$$

Обучение с учителем

- Задано:
  - Предиктор и таргет (x,y) *изображения - метки*
  - Семейство алгоритмов $a_\theta(x) \to y$ линейные/деревья/НС
  - Лосс функция $L(y, a_\theta(x))$ MSE/crossentropy

- Найти:

$$\theta' \leftarrow argmin_\theta L(y, a_\theta(x))$$

Интернет запросы

У нас есть:

- YouTube
- Прямой поток данных

(баннер и видео, #clicked)

Мы хотим:

- Научиться выбирать релевантные запросы

Идеи?

Решение в лоб

- **Общая идея:**
  - Инициализировать с наивным решением
  - Получить данные методом проб и ошибок, и ошибок, и ошибок
  - Изучите (ситуацию) → (оптимальное действие)
  - Повторять

Гигантский робот смерти (ГРС)

- **У нас есть:**
  - Злой человекоподобный робот
  - Много запчастей для его ремонта :)
- **Мы хотим:**
  - ~~Поработить человечество~~
  - Научиться ходить вперед

Решение в лоб

- **Общая идея:**
  - Инициализировать с наивным решением
  - Получить данные методом проб и ошибок, и ошибок, и ошибок
  - Изучите (ситуацию) → (оптимальное действие)
  - Повторять

Проблемы

Проблема 1

- Что именно означает "оптимальное действие"?

Извлечь как можно больше денег, сколько сможете прямо сейчас

VS

Сделать пользователя счастливым чтобы он посетил вас снова

Проблемы

Проблема 2:

- Если вы всегда следуете "текущей оптимальной" стратегии, вы можете никогда не обнаружить что-то лучше.
- Если вы показываете один и тот же баннер 100% пользователей, вы никогда не узнаете, как на них влияют другие объявления.

Идеи?

Reinforcement Learning

Бандит

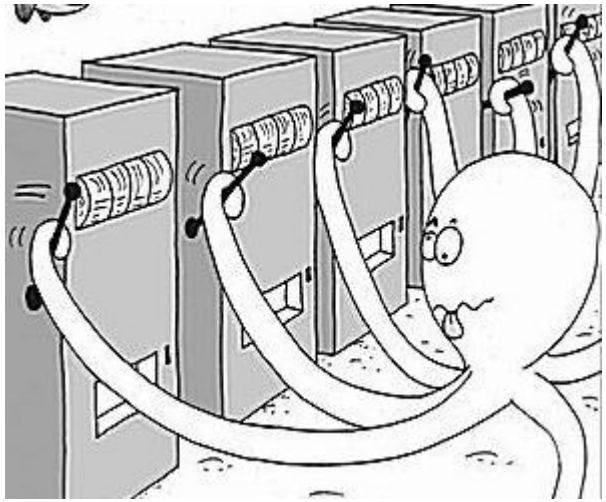
Примеры:

- баннерная реклама (RTB)
- рекомендации
- лечение

Бандит

Примеры:

- баннерная реклама (RTB)
- рекомендации
- лечение



Бандит

Примеры:

- баннерная реклама (RTB)
- рекомендации
- лечение

Вопрос: что такое наблюдение, действие и обратная связь в проблеме баннерной рекламы?

Бандит

Примеры:

- баннерная реклама (RTB)
- рекомендации
- лечение

Вопрос: что такое наблюдение, действие и обратная связь в проблеме баннерной рекламы?

Бандит

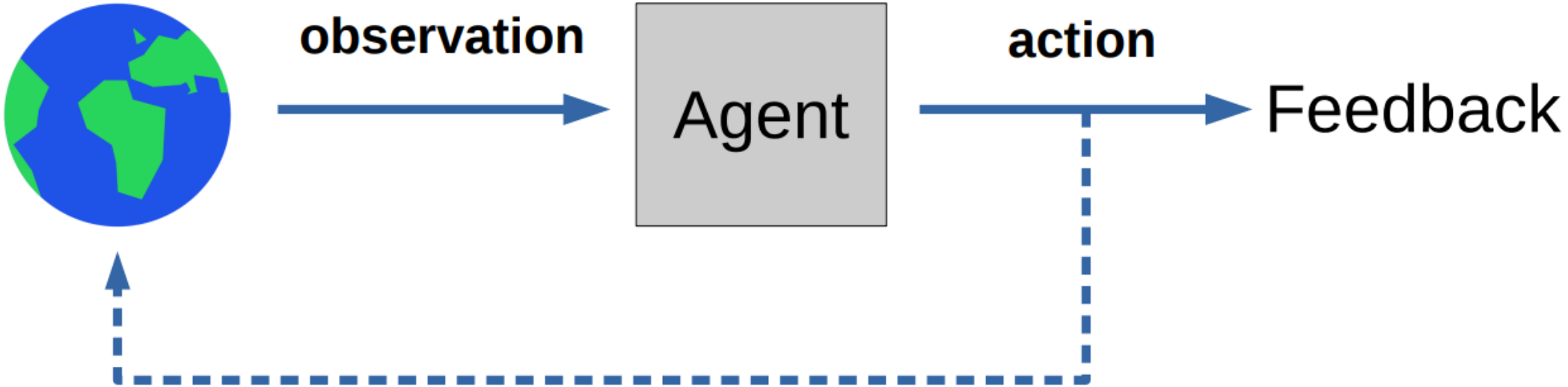
Вопрос: Вы - Яндекс/Google/Youtube.

Есть вид баннеров, которые имеют

большой процент кликов: "кликбейт".

Хорошая ли это идея - показывать кликбейт?

Бандит



Вопрос: Вы - Яндекс/Google/Youtube.

Есть вид баннеров, которые имеют большой процент кликов: "кликбейт".

Хорошая ли это идея - показывать кликбейт?
Нет, после этого вам никто не будет доверять!

Суммарной вознаграждение

- Суммарное вознаграждение за сессию:

$$R = \sum_t r_t$$

- Политика агента
  - ✓ $\pi(a|s) = P[a_t = a|s_t = s]$

- Задача: найти политику с максимальным вознаграждением

$$\pi(a|s): E_\pi[R] \to max$$

Цель

- Простой способ:

$E_\pi R$ - это ожидаемая сумма вознаграждений которую агент с политикой зарабатывает за сессию

- Сложный способ:

$$E_{s_0, r_0 \sim p_0} E_{s_1, r_1 \sim p_1} \dots E_{s_T, r_T \sim P(s', r | s_{T-1}, a_{T-1})} [r_0 + r_1 + \dots + r_T]$$

Как будем действовать

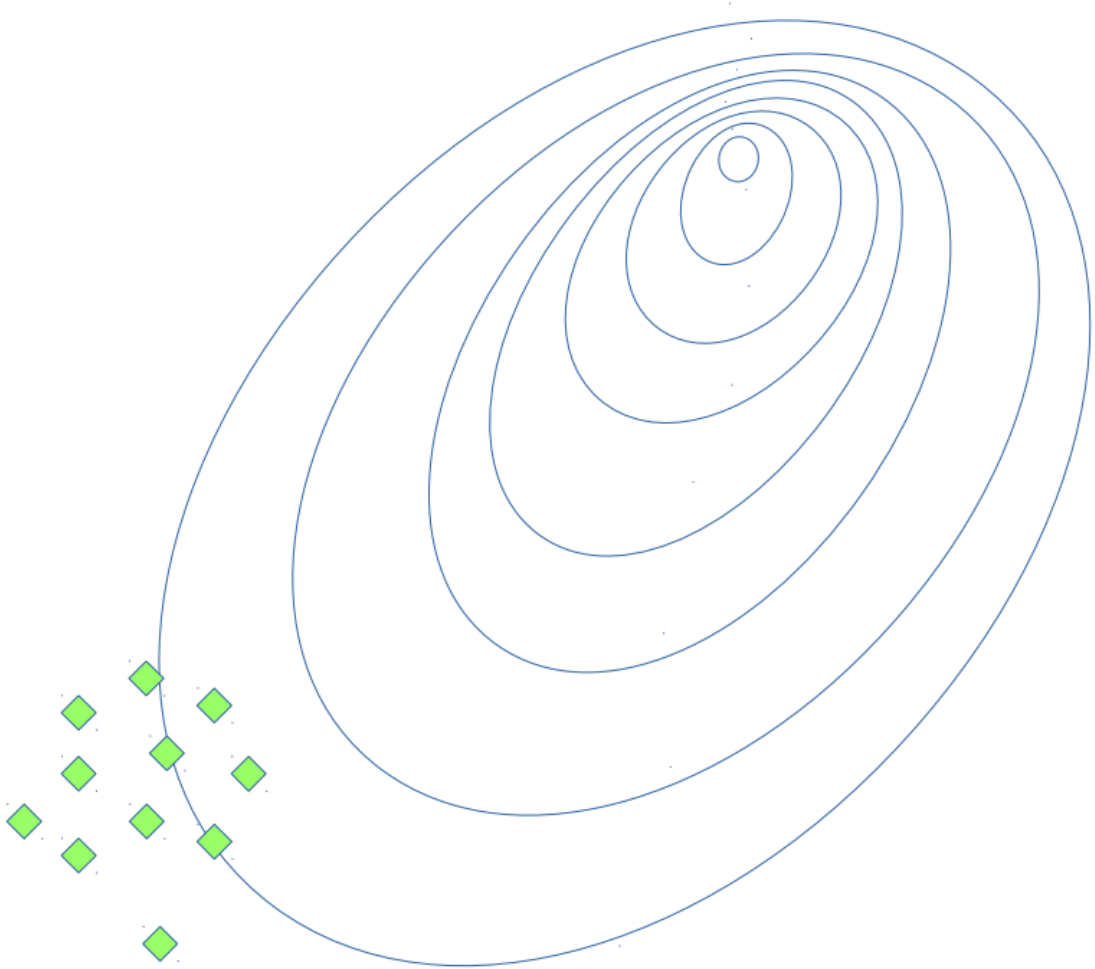
- Общая идея
  - Сыграть несколько сессий
  - Обновить политику
  - Повторить

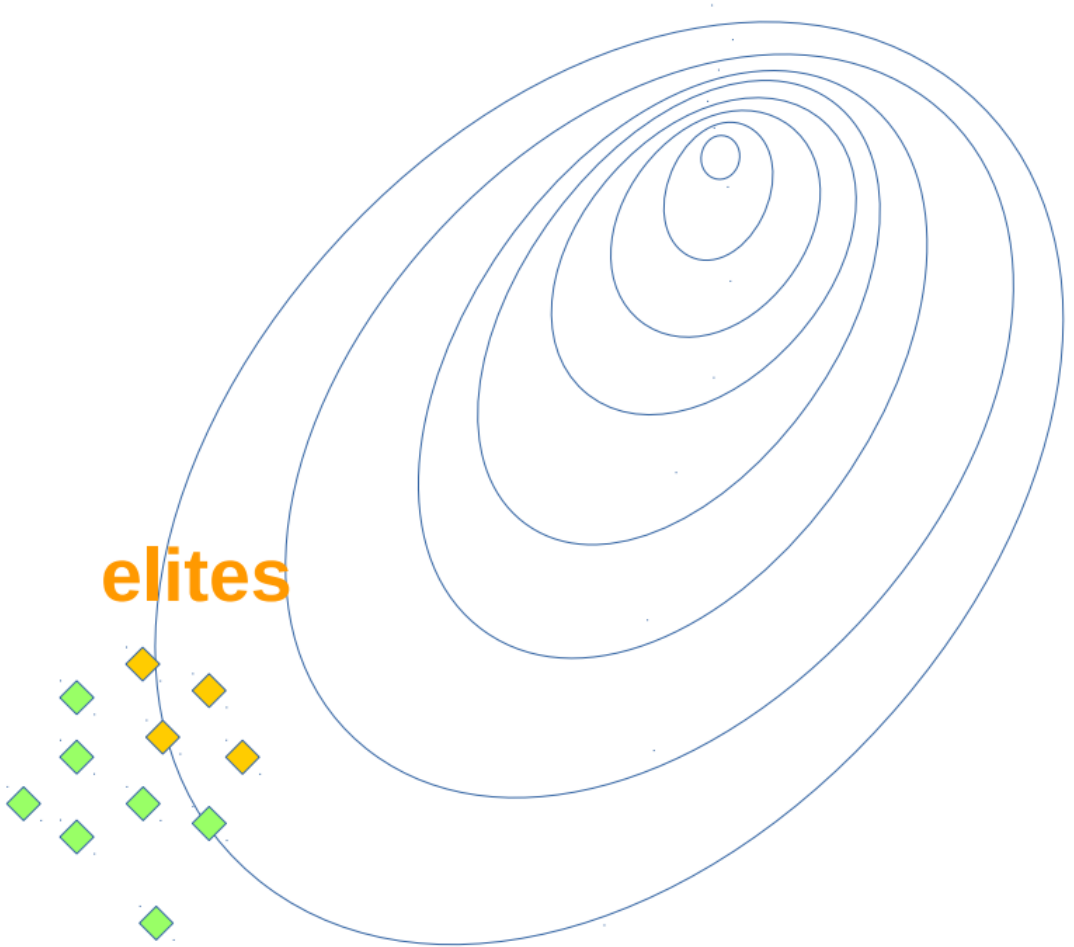
Метод crossentropy

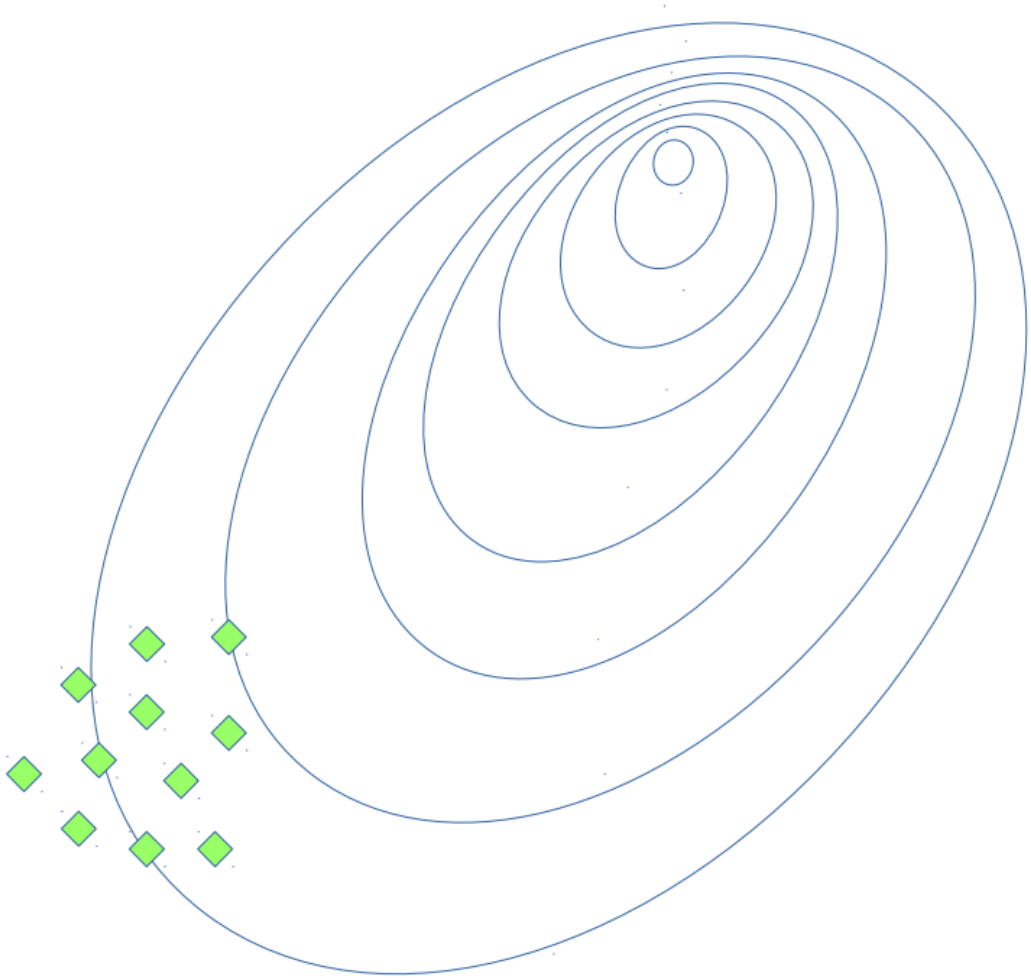
Инициализировать политику

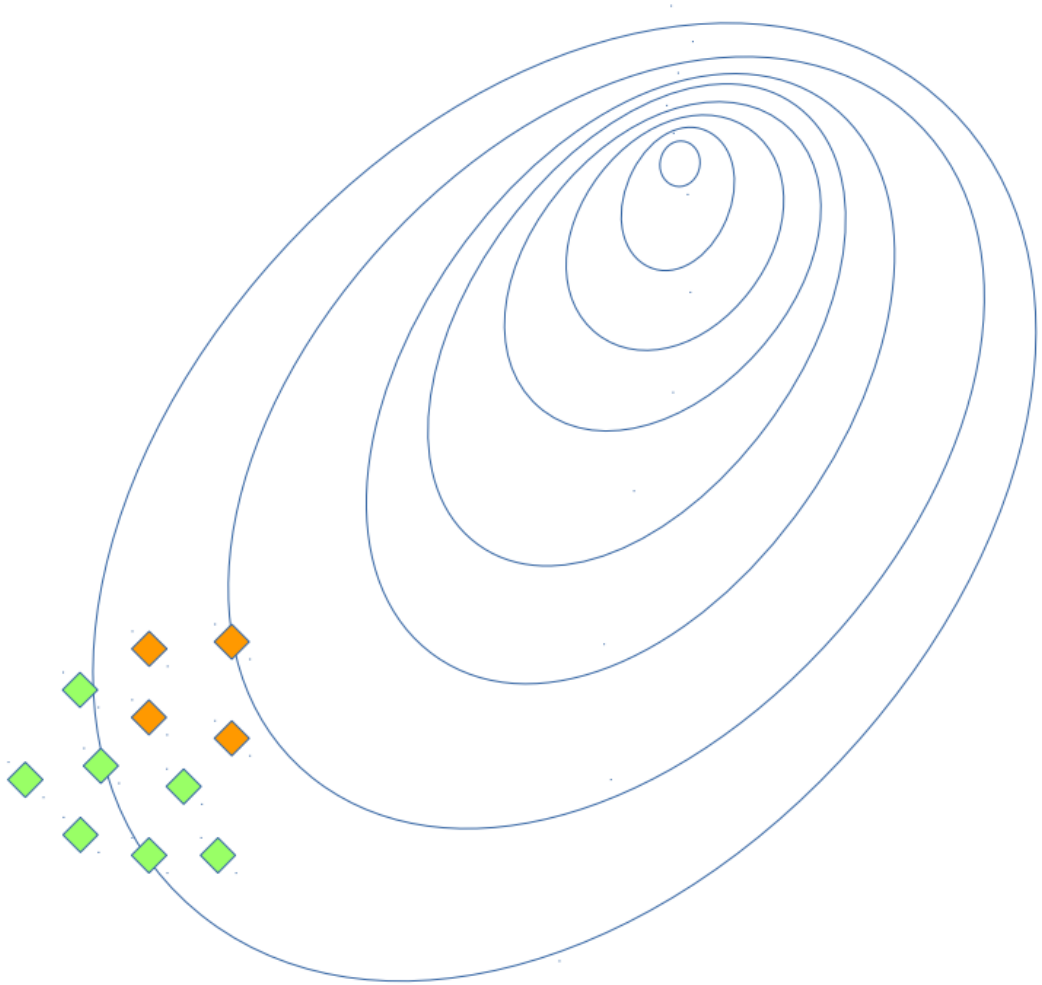
Повторить:

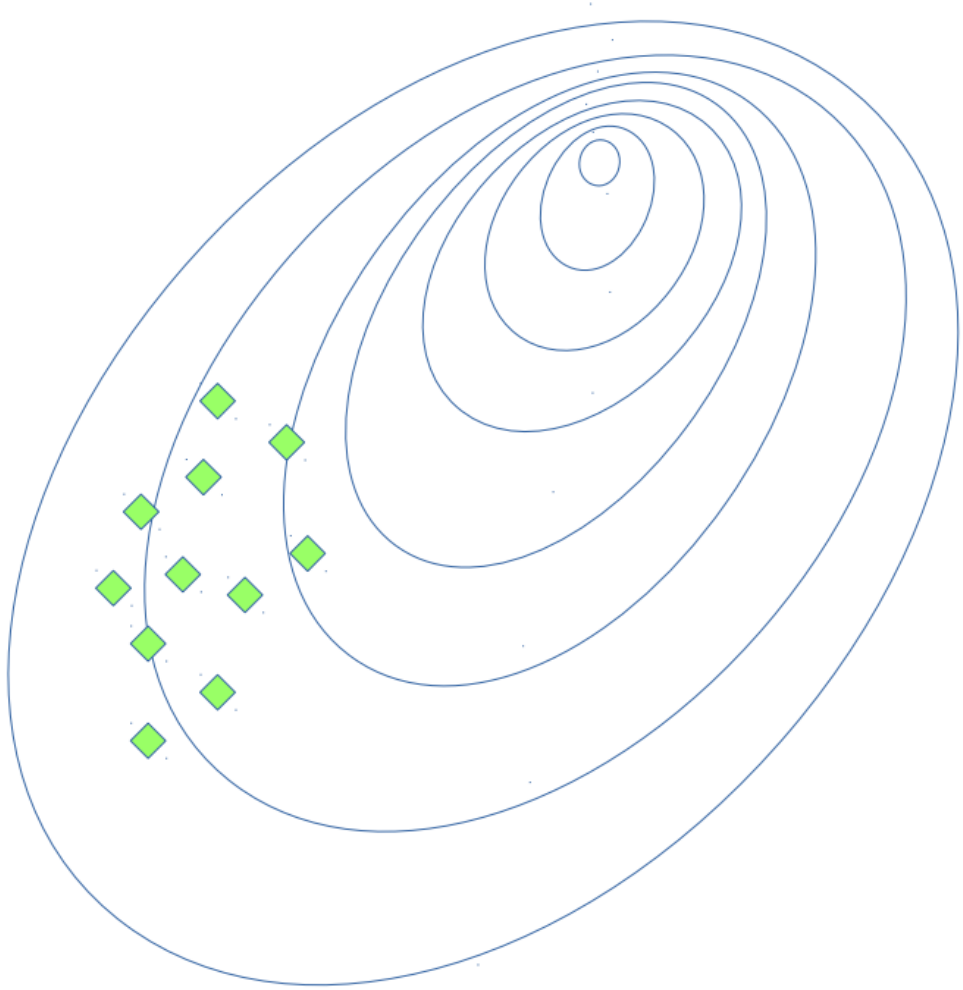
- Сыграть N[100] сессий
- Выбрать M[25] лучших сессий, называемых элитными сессиями
- Изменить политику таким образом, чтобы приоритет отдавался действиям с элитных сессий

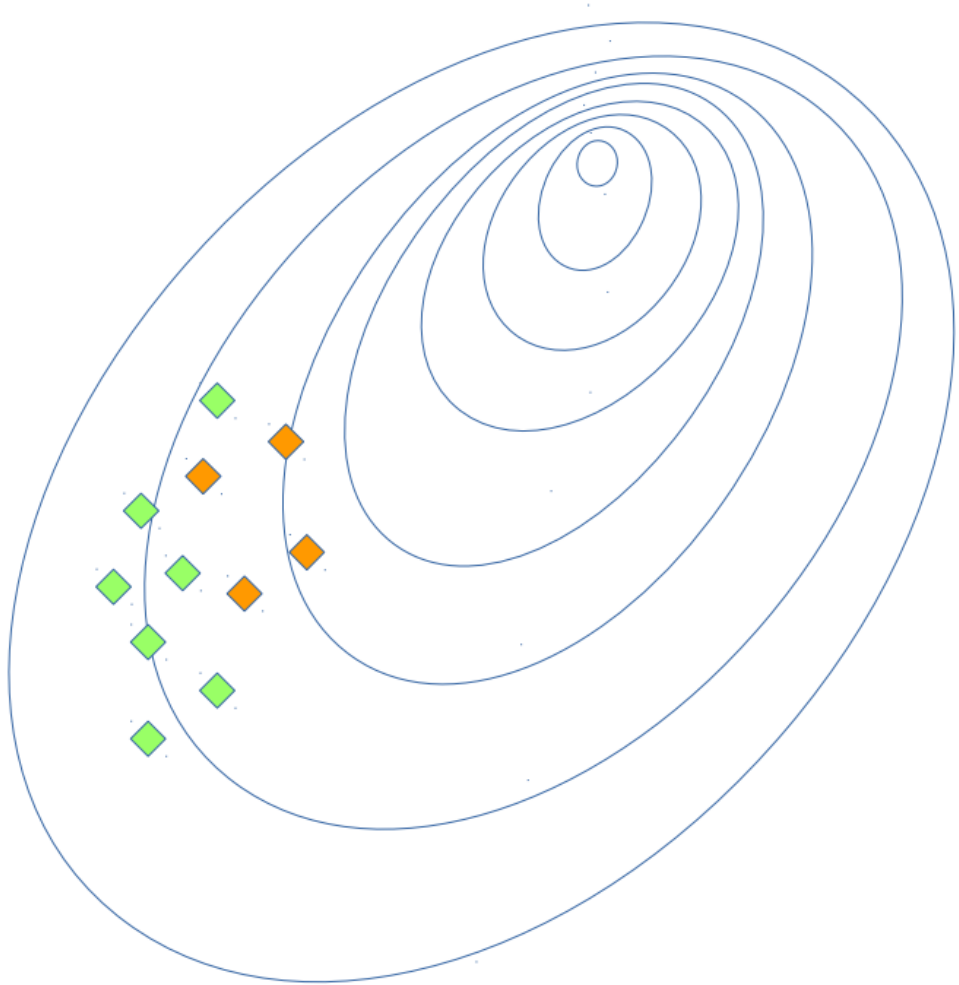


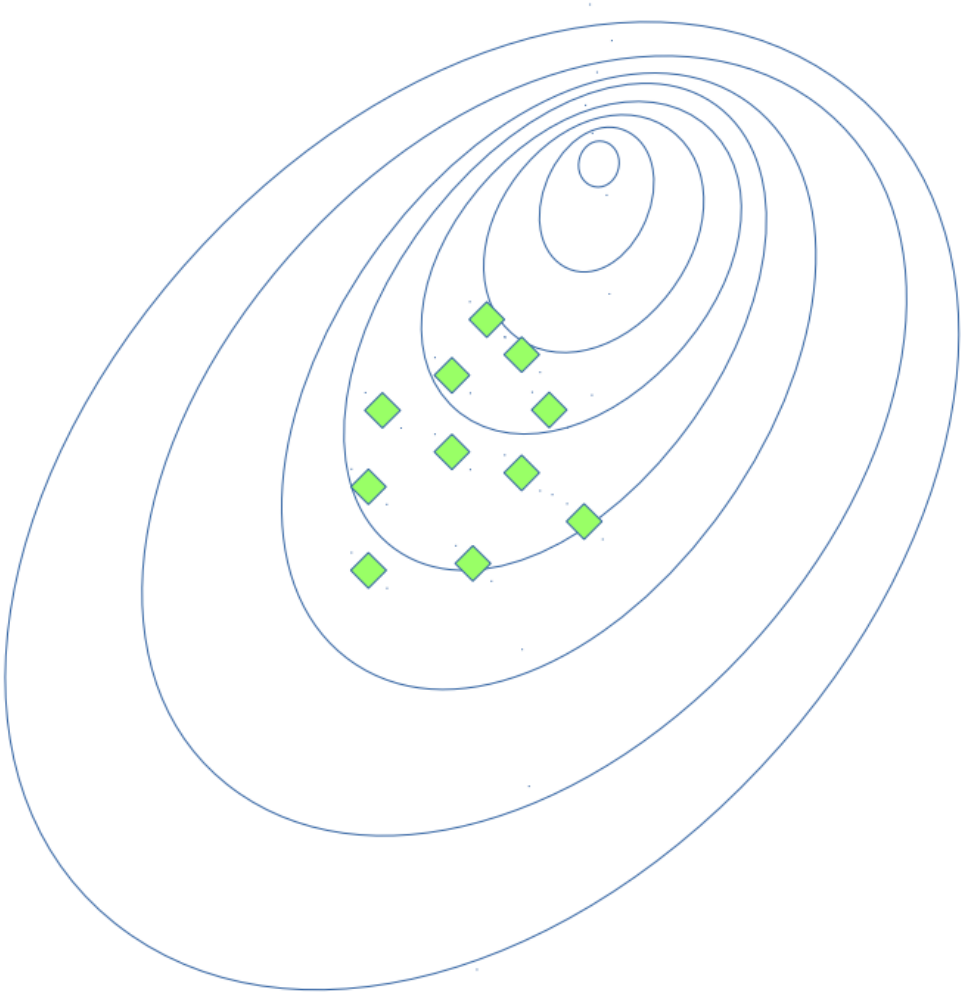

elites

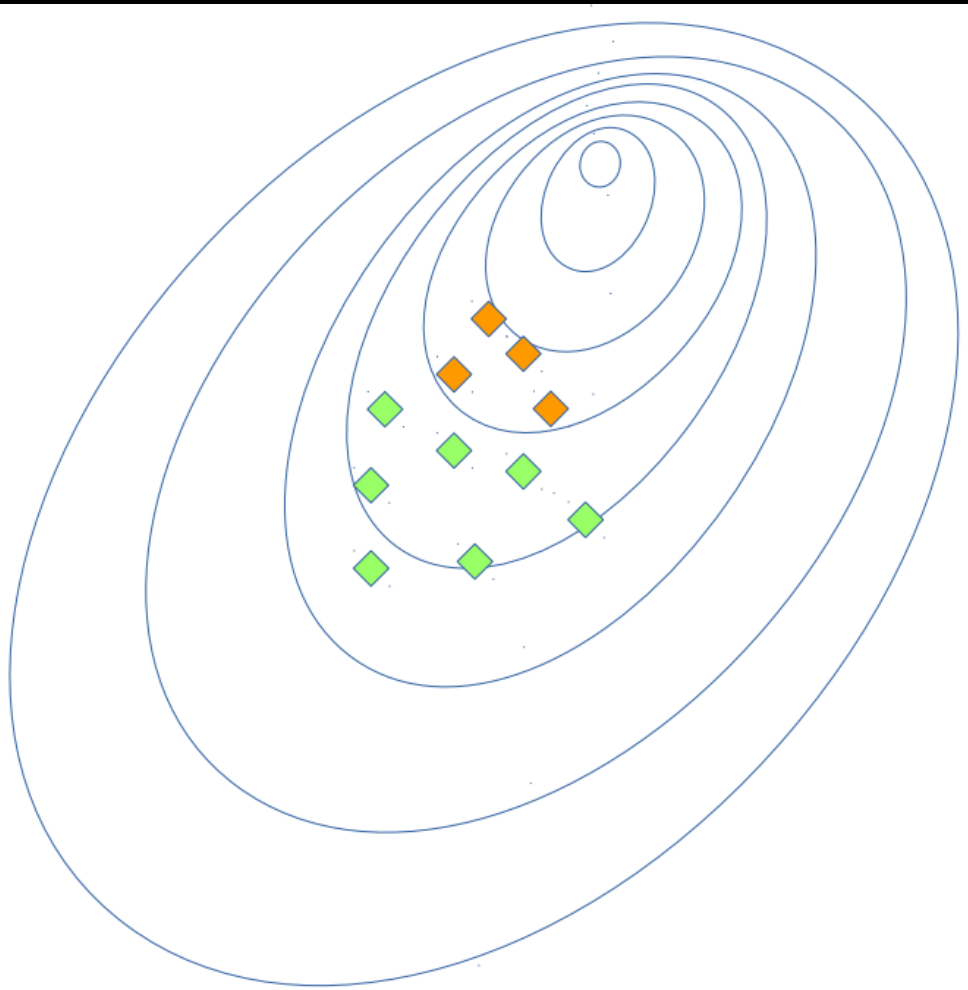


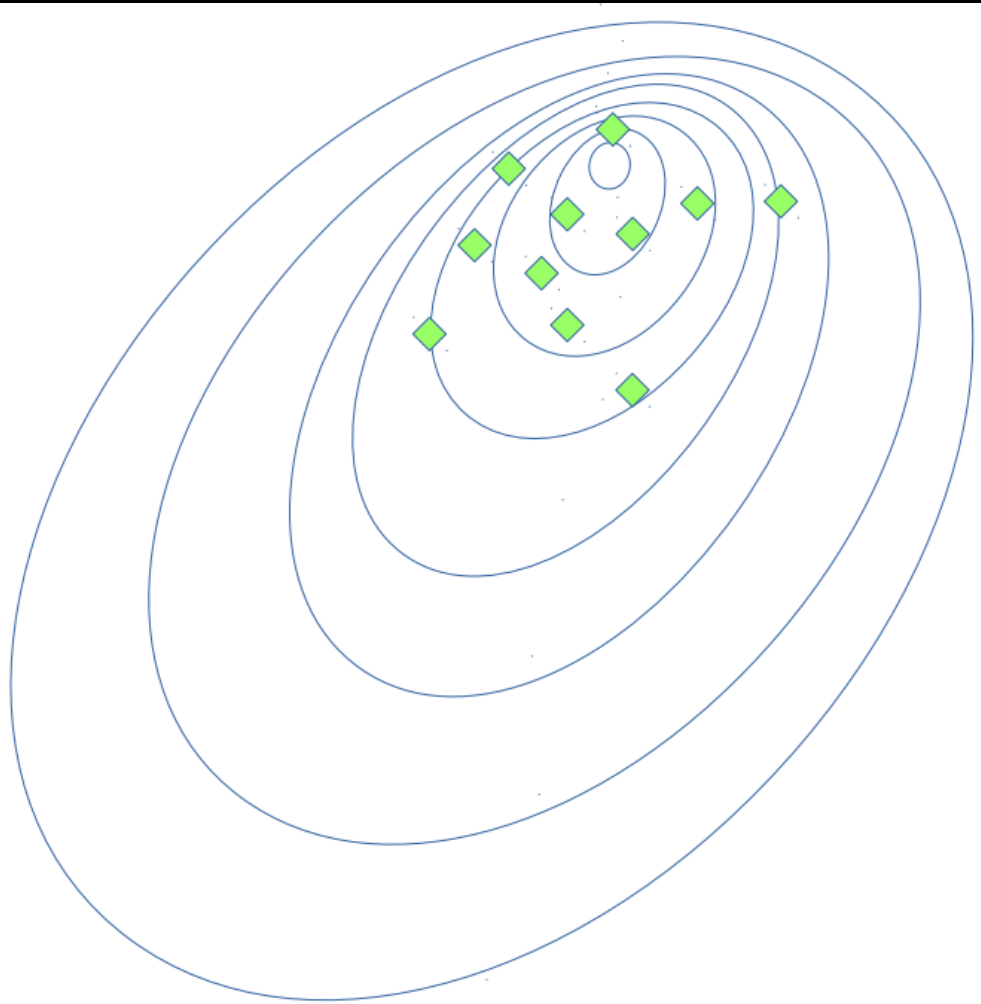


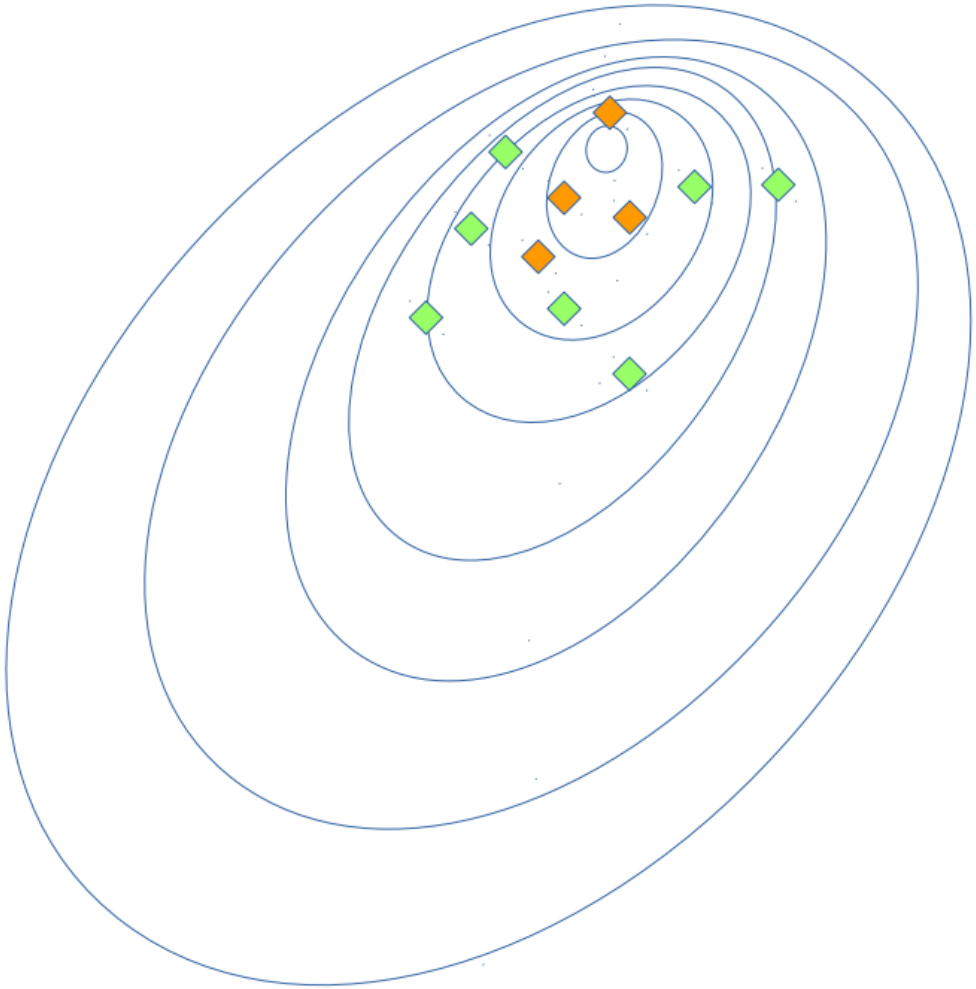


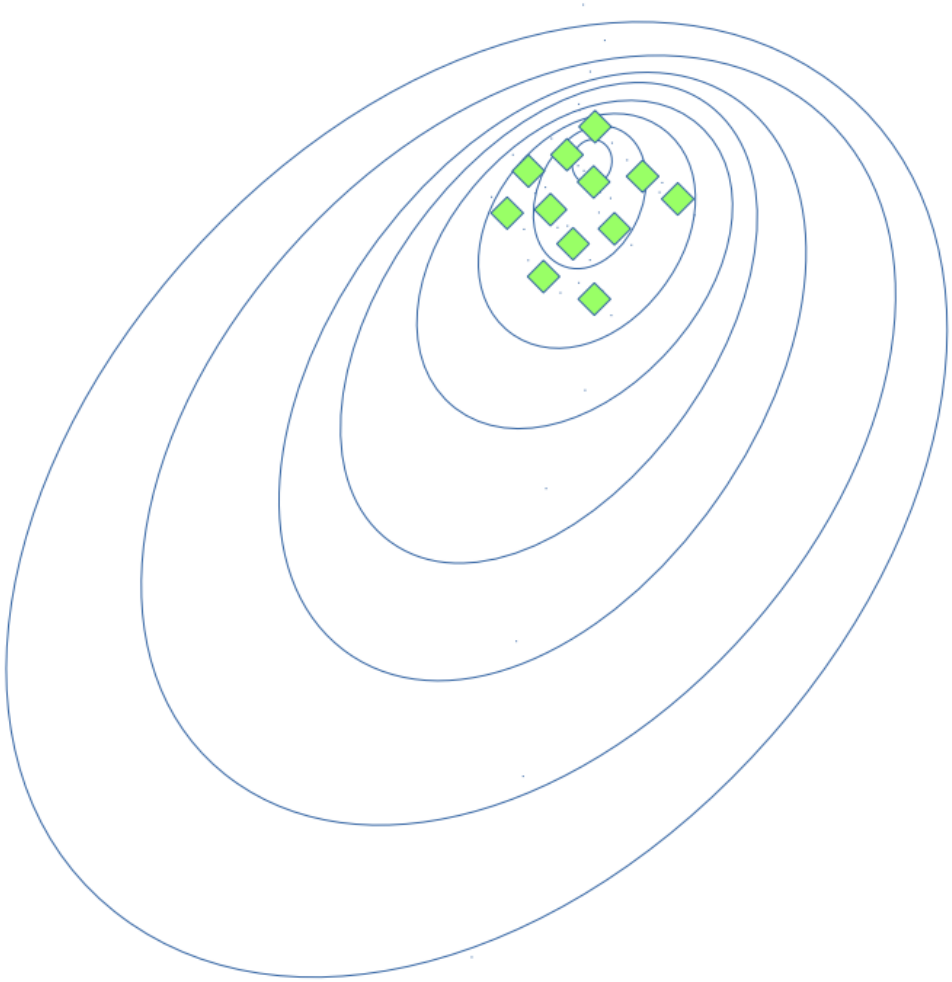












Метод табличной кроссэнтропии

- Политика – это матрица

$$\pi(a|s) = A_{s,a}$$

- Сыграть N игр с этой политикой
- Получить M лучших сессий (элитных)

$$elite = [(s_0, a_0), (s_1, a_1), \dots, (s_k, a_k)]$$

Метод табличной кроссэнтропии

- Политика – это матрица

$$\pi(a|s) = A_{s,a}$$

- Сыграть N игр с этой политикой
- Получить M лучших сессий (элитных)
- Объединить по состояниям

$$\pi(a|s) = \frac{\sum_{s_t,a_t \in Elite}[s_t = s][a_t = a]}{\sum_{s_t,a_t \in Elite}[s_t = s]}$$

Среды с бесконечным/большим пространством состояний

Приближенный метод кроссэнтропии

- Политика аппроксимирована
  - Нейронная сеть предсказывает $\pi(a|s)$ при заданном s
  - Линейная регрессия, деревья решений и т.д.

- Невозможно установить $\pi(a|s)$ в явном виде
- M лучших сессий (элитных)

$$elite = [(s_0, a_0), (s_1, a_1), \dots, (s_k, a_k)]$$

Приближенный метод кроссэнтропии

Нейронная сеть предсказывает $\pi(a|s)$ при заданном s

M лучших сессий (элитных)

$$elite = [(s_0, a_0), (s_1, a_1), \dots, (s_k, a_k)]$$

Максимизируем правдоподобие действий в лучших играх

$$\pi = \underset{\pi}{\operatorname{argmax}} \sum_{s_i, a_i \in Elite} log\pi(a_i|s_i)$$

Приближенный метод кроссэнтропии

- Инициализировать веса

- Цикл:

  Сэмплируем N сессий

  $elite = [(s_0, a_0), (s_1, a_1), \dots, (s_k, a_k)]$

  $$w_{i+1} = w_i + \alpha \nabla \left[\sum_{s_i, a_i \in Elite} \log \pi_{w_i}(a_i | s_i) \right]$$