Reinforcement Learning

Multi Armed Bandits
Contextual Bandits

Александр Костин
telegramm: @Ko3tin
LinkedIn: kostinalexander

Recap

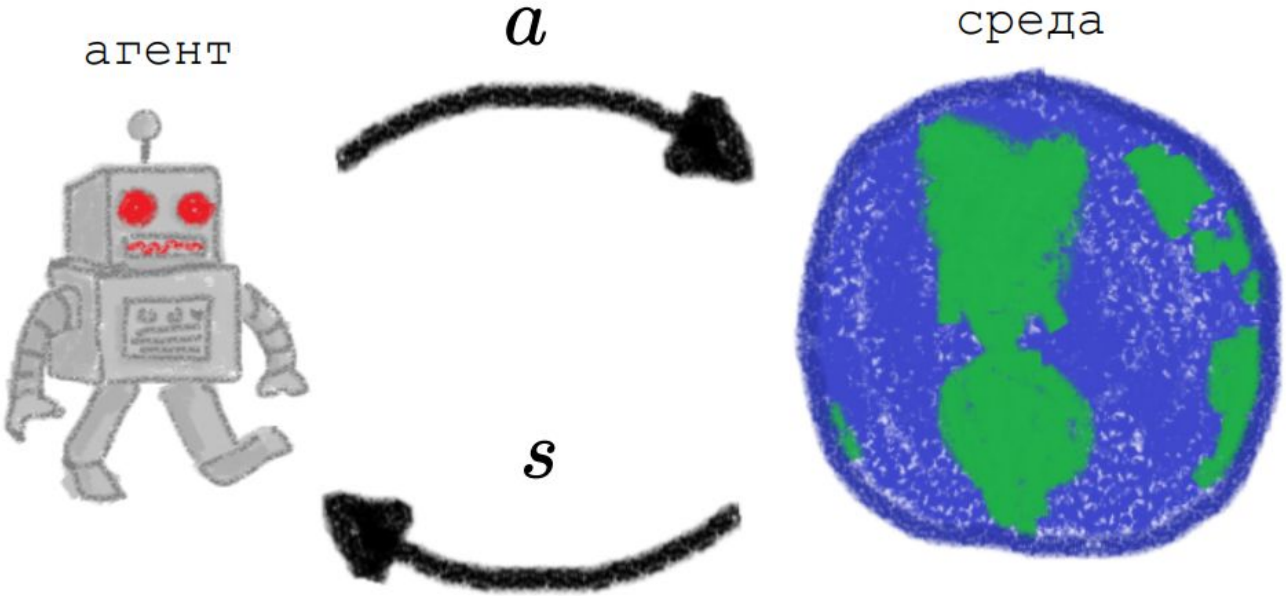
Политика агента:

a = π(a|s)

a ~ π(a|s)

Цель агента:

$$\mathbb{E}_{p(\tau|\pi)}(\sum_{t=0}^{T} r_t) \rightarrow \max_\pi$$



Recap

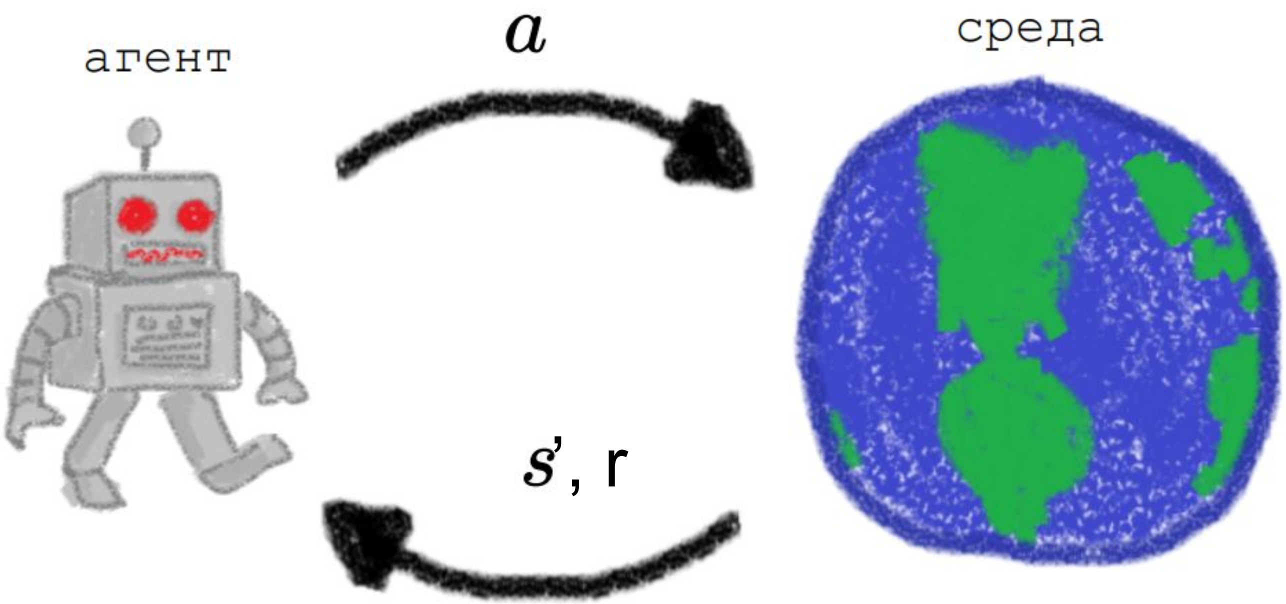
Политика агента:

a = π(a|s)

a ~ π(a|s)

Цель агента:

$$\mathbb{E}_{p(\tau|\pi)}\left(\sum_{t=0}^{T} r_t\right) \to \max_\pi$$



Рекомендация музыки

Имеем:

- Есть много разных пользователей
- Есть много разной музыки

Хотим:

- Рекомендовать музыку
- Пользователи продолжали пользоваться сервисом

What if

- **s’** не зависит от **s** и **a**
- все **s** одинаковы

Получаем многорукого бандита

Многорукие бандиты

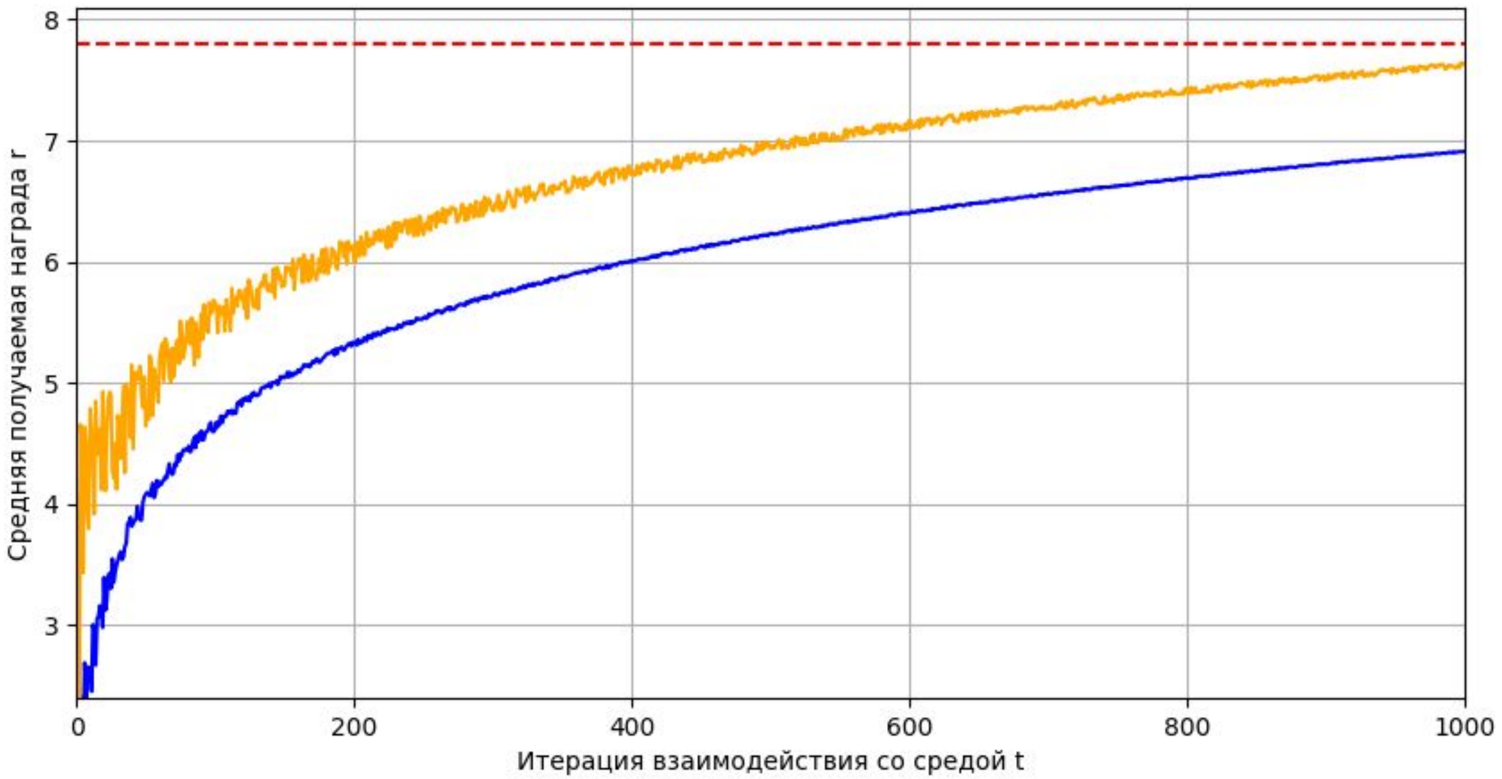
Exploration vs Exploitation Dilemma

- Принятие решений в режиме онлайн предполагает фундаментальный выбор:
    - Exploitation. Использование имеющихся знаний для максимизации награды здесь и сейчас
    - Exploration. Исследовать среду. Собрать больше информации
- Лучшая долгосрочная стратегия может включать краткосрочные потери
- Чтобы добиться наилучшего результата в целом решения, необходимо собрать достаточно информации

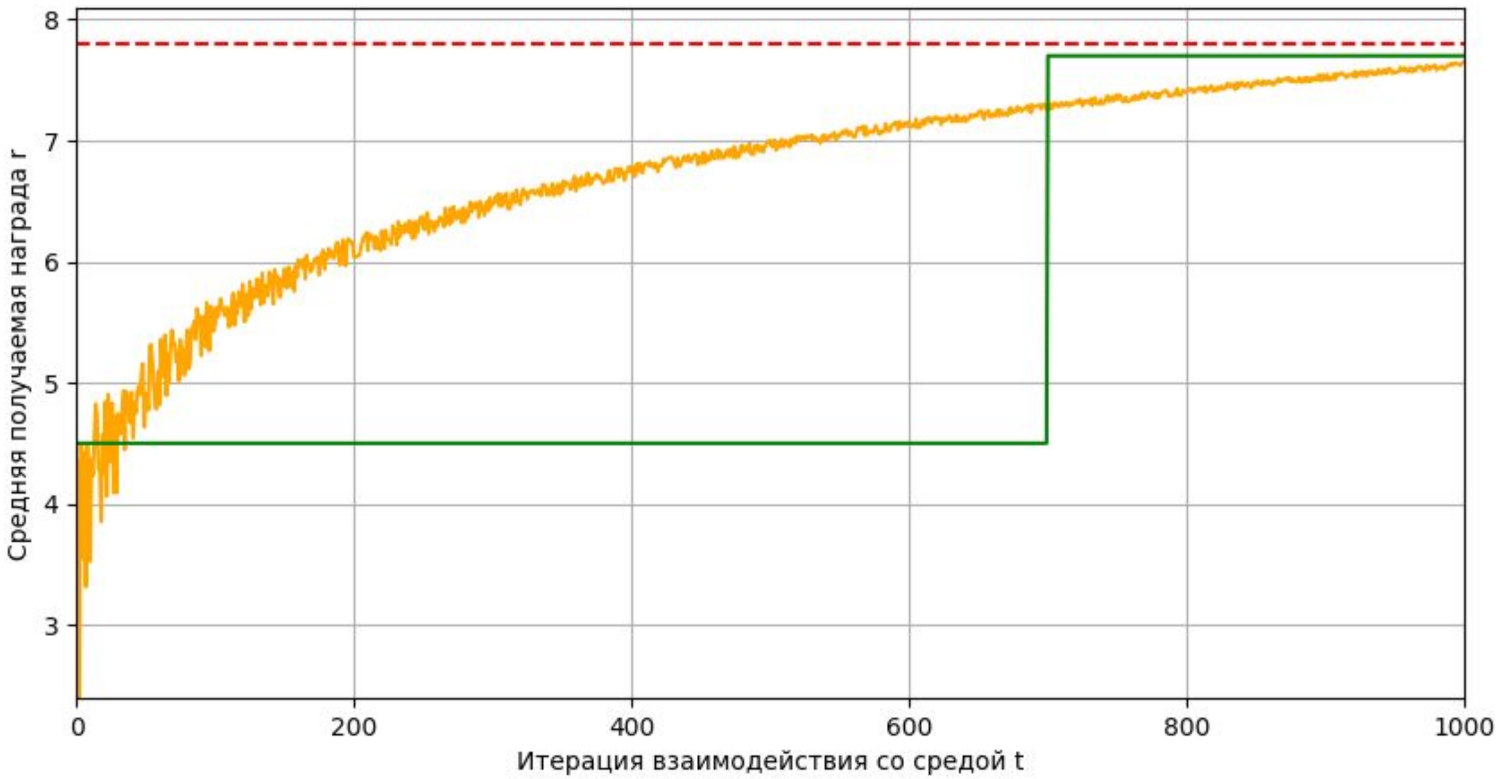
Какая стратегия исследования лучше

Идеи?

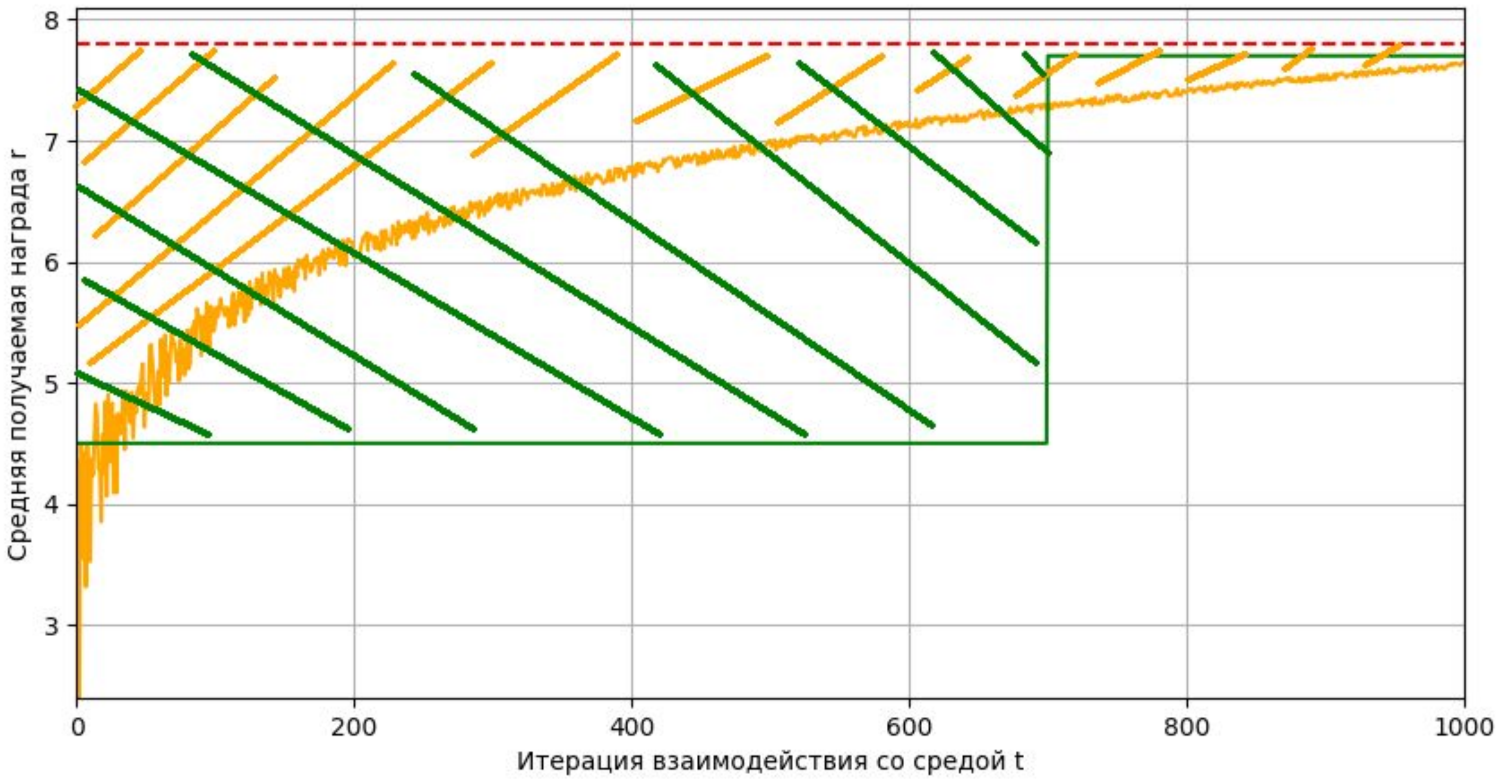
Какая стратегия исследования лучше



Какая стратегия исследования лучше



Какая стратегия исследования лучше



Постановка задачи многорукого бандита

Предположим, что эпизод бесконечен, но в среде имеется только одно состояние. Поэтому связи между последовательными действиями нет. Агент постоянно сталкивается с выбором между различными действиями.

Постановка задачи многорукого бандита

Предположим, что эпизод бесконечен, но в среде имеется только одно состояние. Поэтому связи между последовательными действиями нет. Агент постоянно сталкивается с выбором между различными действиями.

Многорукий бандит – это пара <**R**, **A**>:

- $\{\mathbf{r_a} \in \mathbf{R} \mid \mathbf{a} \in \mathbf{A}\}$ - набор распределений вознаграждений
- На каждом шаге **t** агент выбирает $\mathbf{a_t}$ и получает вознаграждение $\mathbf{r_a}$

Задача агента максимизировать награду:

$$R = E_\pi\left[\sum_{t=1}^{T} r(a_t)\right]$$

Постановка задачи многорукого бандита

Исследование: найти действие, дающее наибольшую награду

Ценность действия: $Q(a) = E[r_t \mid a_t = a]$

Оптимальная награда: $V^\star = \max_a Q(a)$

Regret: $E_\pi[V^\star - Q(a)] \geq 0$

Постановка задачи многорукого бандита

Исследование: найти действие, дающее наибольшую награду

Ценность действия: $Q(a) = E[r_t \mid a_t = a]$

Оптимальная награда: $V^\star = \max_a Q(a)$

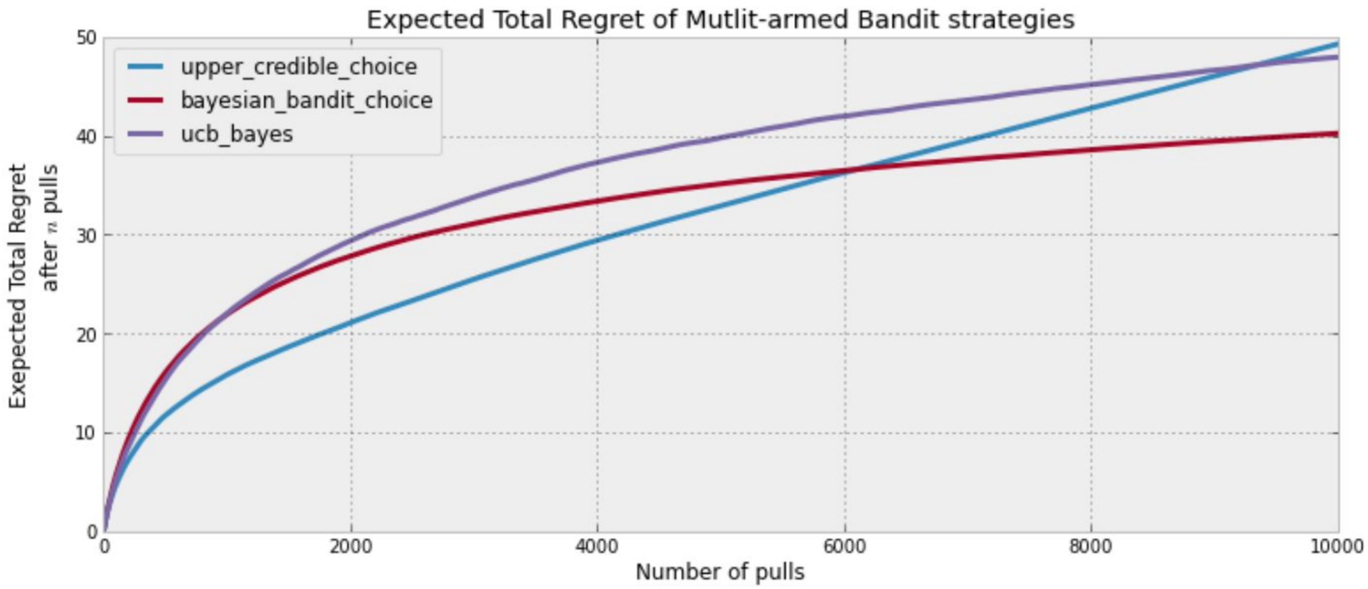
Regret: $E_\pi[V^\star - Q(a)] \geq 0$

Total Regret: $E_\pi \sum_{t=1}^{T} [V^\star - Q(a)] \to \min_\pi \iff E_\pi \sum_{t=1}^{T} [r_t] \to \max_\pi$

Regret

Regret - недополученная награда

$$\eta = E_{\pi}[V^{\star} - Q(a)]$$



ε-greedy Policy

Идея: давайте брать жадное действие с вероятностью **(1 - *ε*)**, и произвольное с вероятностью ***ε***

$$Q(a) = \frac{\Sigma r_a}{N_a}$$

$$\pi(a) = \begin{cases} \arg\max(Q(a)) & if\ p < 1 - \epsilon \\ \frac{\epsilon}{|A|} & otherwise \end{cases}$$

- Такой метод всегда продолжает изучение среды
- Линейный regret при T->inf

Оптимальность исследования

А линейный regret - это плохо или хорошо?

Оптимальность исследования

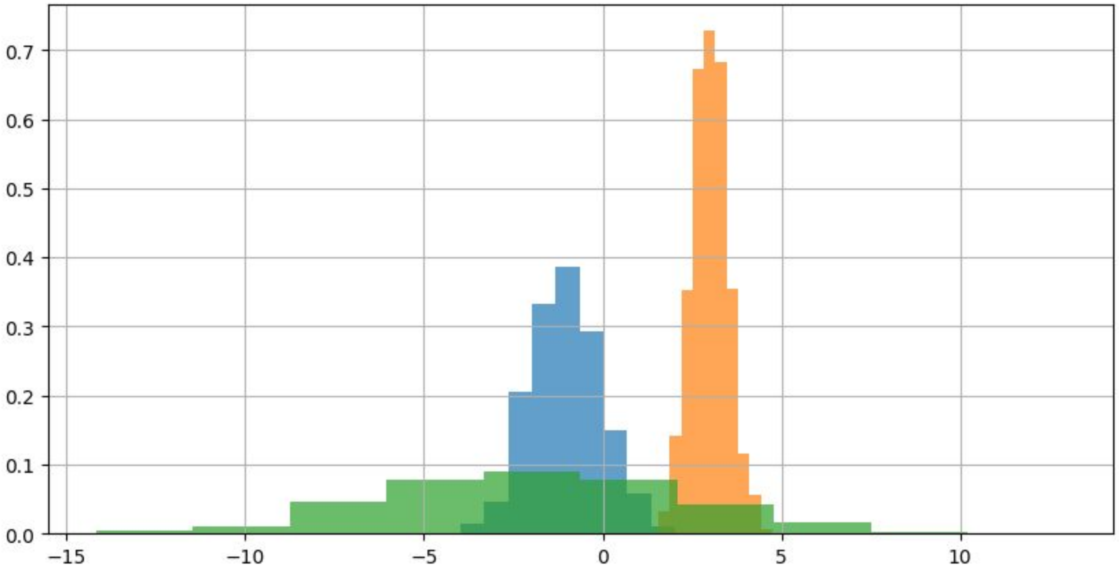
А линейный regret - это плохо или хорошо?

Теорема (Lai and Robbins):

$$\lim_{t\to\infty} Regret \geq \log t \sum_{a \mid \eta_a > 0} \frac{\eta_a}{KL(r_a || r_{a^\star})}$$

ε-greedy Policy

Очевидно, что нет особого смысла дергать синюю ручку на шаге исследования.



Thompson Sampling

1. Хотим пробовать действия пропорционально вероятностям, с которыми они дадут наибольшую награду
2. Зададим априорное распределение на ручках бандита
3. Дергаем ручку бандита с вероятностями:

$$\pi(a \mid h_t) = P\big[Q(a) > Q\big(a'\big),\ a \neq a' \mid h\big]$$

4. Считаем апостериорное распределение по формуле Байеса
5. Возвращаемся к шагу 2

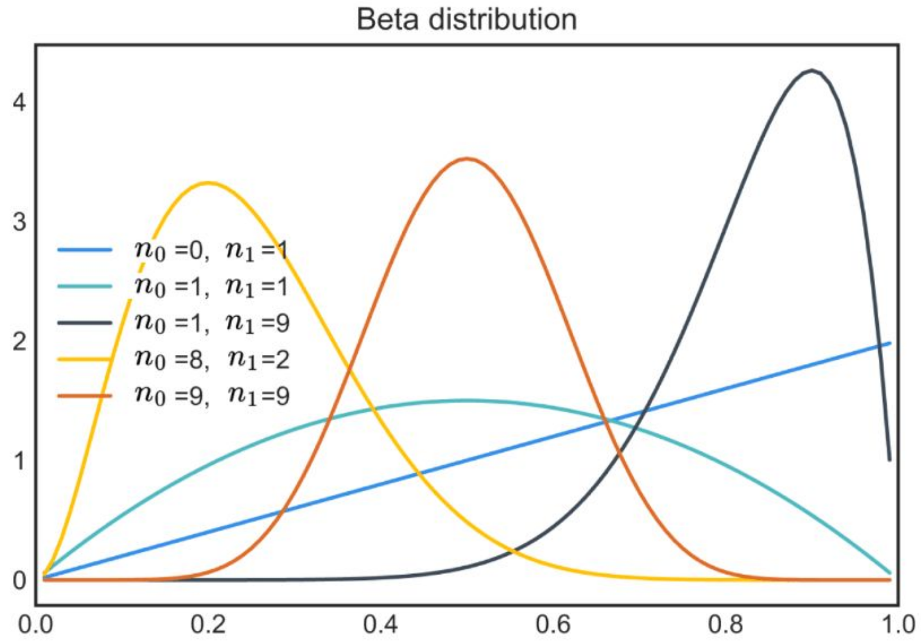
У этого метода логарифмический regret

Thompson Sampling

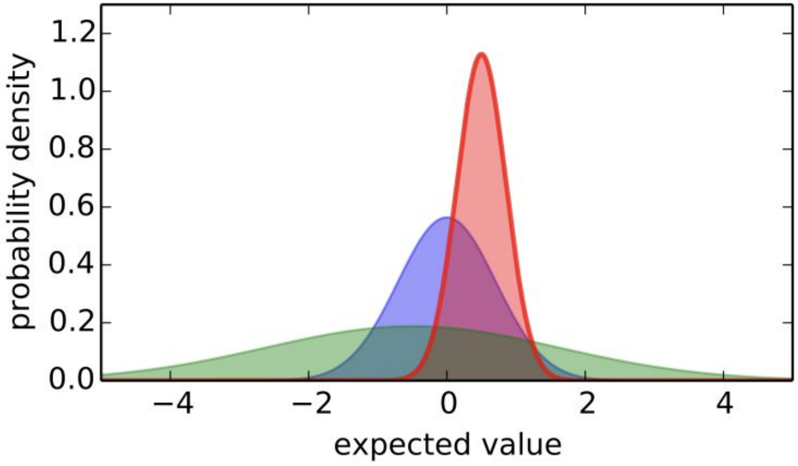
Случай бернуллиевского бандита:

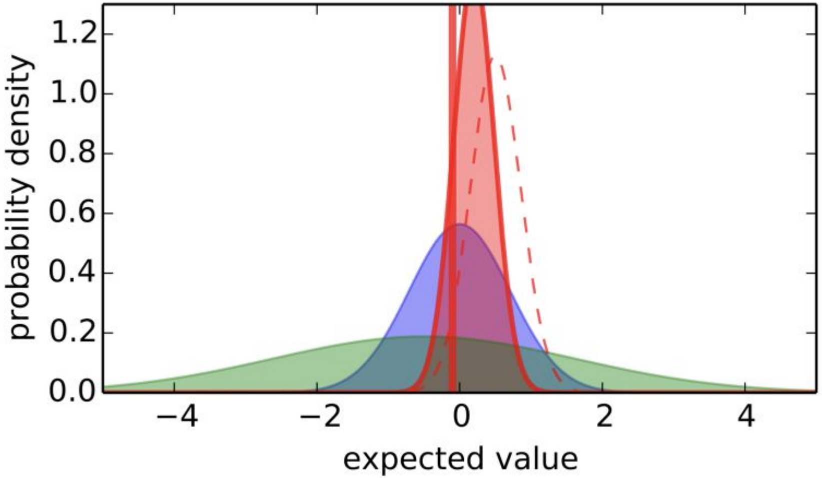
- $r_a = \begin{cases} 1 & if\ p > \theta_a \\ 0 & otherwise \end{cases}$
- $E[Q(a)] = \theta_a$
- Априори: $\theta \sim I[0, 1]$
- Апостериори:

$$P[Q(a)] = Beta(n_1; n_0)$$



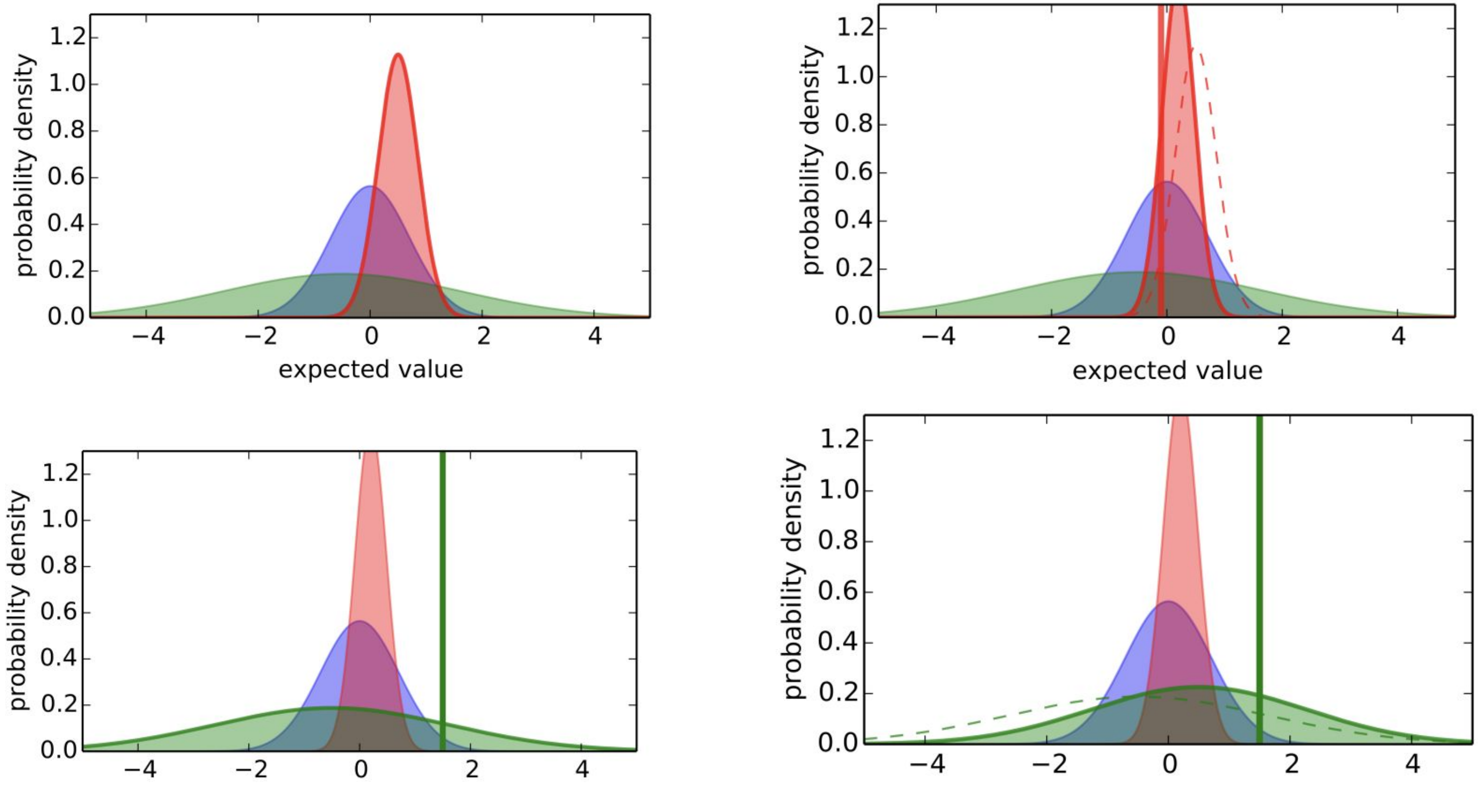
Optimism in face of uncertainty





- Какое действие выбрать?
- Чем более неуверенны мы в отношении ценности действия, тем важнее изучить это действие.
- Оно может оказаться лучшим действием

Optimism in face of uncertainty



Upper Confidence Bound

- Для каждого действия **a** считаем верхнюю границу доверительного интервала **U(a)**
- Ширина интервала зависит от **N(a)** количества использований действия **a**
  - Действие почти не использовалось -> верхняя граница **U(a)** будет огромной. (Неуверенная оценка среднего)
  - Действие использовалось часто -> верхняя граница **U(a)** будет маленькой. (Уверенная оценка среднего)
- В качестве действия выбирается:

$$a_t = \arg\max_{a \in A}[Q_t(a) + U_t(a)]$$

Как оценить верхнюю границу?

Неравенство Хёфдинга для произвольного распределения награды **r**, но **r∈[0,1]**:

- $P[Q_t(a) + U_t(a) < Q(a)] \leq e^{-2N_t(a)U_t(a)^2}$

- Пусть $U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}}$ тогда $e^{-2N_t(a)U_t(a)^2} = p$
- Возьмем $p = \frac{1}{t}$ тогда $\boxed{U_t(a) = \sqrt{\frac{\log t}{2N_t(a)}}}$

UCB

1. Инициализируем распределения оценки среднего (например $Q_t(a) \sim N(0,1)$)
2. Выбираем действие $$a_t = \arg\max_{a \in A} \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$$
3. Обновляем оценку среднего $\mathbf{Q_t(a)}$ и счетчики **t** и $\mathbf{N_t(a)}$
4. Повторяем с шага 2

При c=sqrt(2) у этого метода логарифмический regret [(Auer et al., 2002)]

Контекстные бандиты

- **s’** не зависит от **s** и **a**
- состояния **s** бывают разными
- награда ручки **a** зависит от состояния **s**

Получаем контекстного многорукого бандита

Рекомендация музыки

Имеем:

- Есть много разных пользователей
- Есть много разной музыки

Хотим:

- Рекомендовать музыку
- Пользователи продолжали пользоваться сервисом

Постановка задачи контекстного бандита

Исследование: найти действие, дающее наибольшую награду

Ценность действия: $Q(s,a) = E[r_t \mid s_t = s;\ a_t = a]$

Оптимальная награда: $V^\star(s) = \max_a Q(s,a)$

Regret: $E_\pi[V^\star(s) - Q(s,a)] \geq 0$

Total Regret: $E_\pi \sum_{t=1}^{T} [V^\star(s) - Q(s,a)] \to \min_\pi \iff E_\pi \sum_{t=1}^{T} [r_t] \to \max_\pi$

LinUCB

- Давайте объединим состояние **s** и действие **a** в виде вектора контекста **c**
- И предположим, что ценность действия **a** линейно зависит от **c**:

$$Q(s_t, a_t) = c_t^T \theta^\star + \epsilon_t; \quad \text{где } \epsilon_t \sim N(0,1)$$

- Минимизация regret'а принимает вид:

$$\max_\pi \left(E \sum_{t=1}^{T} r_t \right) \Leftrightarrow \min_\pi E \left[\sum_{t=1}^{T} \max_{a \in A} \langle c^T ; \theta^\star \rangle - \sum_{t=1}^{T} Q_t \right]$$

$$\Leftrightarrow \min_\pi E \left[\sum_{t=1}^{T} \max_{a \in A} \langle c^T - c_t^T ; \theta^\star \rangle \right]$$

Линейная регрессия

Предположим на шаге t имеем несколько пар **{c; Q}** - наш датасет для обучения линейной регрессии:

$$\hat{\theta}_t = \arg\min_{\theta \in R^d} \sum_{k=1}^{t-1} \left(Q_k - c_k^T \theta\right)^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

Решение:

$$\hat{\theta}_t = V_{t-1}^{-1} \sum_{k=1}^{t-1} c_k Q_k$$

$$\text{где } V_{t-1} = \sum_{k=1}^{t-1} c_k c_k^T + \lambda I_d$$

Optimism in face of uncertainty

- В обычном UCB мы искали верхнюю грань доверительного интервала

- В линейной модели мы ищем вектор параметров θ в эллипсоиде $M \in R^d$:

$$a_t = \arg\max_{a \in A} c^T \hat{\theta}_t \Rightarrow a_t = \arg\max_{a \in A} \max_{\theta \in M_t} c^T \theta$$

Confidence Ellipsoid

Пусть $$\beta_t(\delta) = \lambda + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t}{\lambda d}\right)}$$

тогда $$M_t(\delta) = \left\{\theta \in R^d : \left\|\theta^\star - \hat{\theta}_t\right\|_{V_{t-1}} \leq \beta_{t-1}(\delta)\right\}$$

задает доверительный эллипсоид для θ^* с уровнем уверенности 1-δ

LinUCB

- На входе: вероятность δ, размерность d, регуляризация λ
- Инициализация: $b = 0_{R^d}; \quad V = \lambda I; \quad \hat{\theta} = 0_{R^d}$
- for t > 0:
  - Получаем состояние **s**
  - Вычисляем: $\beta_{t-1}(\delta) = \lambda + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{t-1}{\lambda d}\right)}$
  - for a in A:
    - Собираем вектор **c**
    - Вычисляем: $UCB(a) = c^T\hat{\theta} + \beta_{t-1}\sqrt{c^T V^{-1} c}$
  - $a_t = \arg\max UCB(a)$
  - Обновляем: $V = V + c_t c_t^T$

$$b = b + r_t c_t$$

$$\hat{\theta} = V^{-1} b$$

Regret порядка d√2

Полезные ссылки

1) D-LinUCB - бандит для динамической среды

2) Neural Contextual Bandits - линейные бандиты с нейронками для исправления нелинейности

3) Лекция по LinUCB