Reinforcement Learning

Lecture 10:
Multi-Arm Bandits

Yishay Mansour, Tel-Aviv University

Lecture 10: outline

- Multi-Arm bandits
- Regret minimization
  - Successive Elimination
  - Upper Confidence Bound
    - UCB
- Best Arm Identification
  - PAC
  - Median Elimination
- Applications

Multi-Arm Bandits

- Model
  - Single state MDP
- Planning
  - Trivial
  - Select highest reward action
- Learning
  - Exploration Exploitation tradeoff
  - Best Arm Identification
  - Cumulative reward

Applications of MAB

- News
  - News article recommendations
- Medical trials
  - Multiple treatments
- Ad selection
  - Multiple advertiser’s ads
- A/B testing
  - dynamic

MAB model

- States:
  - One state
  - No dynamics
- Actions
  - $K = \{1, \dots, k\}$
- Rewards
  - Action i has reward X_i
    - Drawn from D_i
    - $\mu_i = E[X_i]$
  - Bounded rewards
    - $X_i \in [0,1]$
  - Best action
    - $\mu^* = \max_i \mu_i$
    - $a^* = \arg\max_i \mu_i$

Cumulative Reward

- Finite horizon
  - T time steps
- At time t
  - Reward action a_t
    - $r_t(a_t)$
- Cumulative reward
  - $\sum_{t=1}^{T} r_t(a_t)$
- Regret
  - $E[\max_i \sum_{t=1}^{T} r_t(i) - \sum_{t=1}^{T} r_t(a_t)]$
- Pseudo regret
  - $\max_i E[\sum_{t=1}^{T} r_t(i)] - E[\sum_{t=1}^{T} r_t(a_t)]$
  - $= \mu^* T - E[\sum_{t=1}^{T} \mu_{a_t}]$

Chernoff-Hoeffding Inequality

- Given m i.i.d. random variables $X_i \in [-1,1]$
  - $\mu = E[X_i]$

- Then
  - $\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \epsilon\right] \leq 2\exp(-\frac{\epsilon^2}{2}m)$

Warmup: Full information $k=2$

- At time t
  - Select action greedy
  - $a_t = \text{argmax}_i \, avg_t(i)$
  - Get reward $r_t(a_t)$
  - Observe $(r_t(1), r_t(2))$
  - Set
    - $avg_t(i) = \frac{1}{t}\sum_{\tau=1}^{t} r_\tau(i)$

Full information $K=2$: Analysis

- Assume $\mu_1 \geq \mu_2$
  - $\Delta = \mu_1 - \mu_2 \geq 0$
- Pseudo Regret
  - $\sum_{t=1}^{\infty}(\mu_1 - \mu_2)\Pr[avg_t(2) \geq avg_t(1)]$
  - $E[avg_t(2) - avg_t(1)] = \mu_2 - \mu_1 = -\Delta$
  - $\Pr[avg_t(2) - avg_t(1) + \Delta \geq \Delta] \leq e^{-\frac{\Delta^2 t}{2}}$
    - Chernoff-Hoeffding

Full information $K=2$: Regret

- $E[Pseudo\ Regret]$
  - $= \sum_{t=1}^{\infty} \Delta \Pr[avg_t(2) \geq avg_t(1)]$
  - $\leq \sum_{t=1}^{\infty} \Delta\, e^{-\Delta^2 t/2}$
  - $\leq \int_0^{\infty} \Delta\, e^{-\frac{\Delta^2 t}{2}} dt$
  - $= \left[\frac{2}{\Delta} e^{-\frac{\Delta^2 t}{2}}\right]_0^{\infty} = \frac{2}{\Delta}$
- Bound independent of T !

Multi-Arm Bandits

- We observe only the payoff of the selected action
  - Time t
  - Select a_t
  - Observe $r_t(a_t)$

- Need to consider exploration.
- Can we have a regret independent of T?

MAB: lower bound

- Consider two equally likely profiles
  - $a_1 \sim \mathrm{Br}\left(\frac{1}{2}\right)$ and $a_2 \sim \mathrm{Br}(\frac{1}{4})$
  - $a_1 \sim \mathrm{Br}\left(\frac{1}{2}\right)$ and $a_2 \sim \mathrm{Br}(\frac{3}{4})$
- Assume that
  - $E[regret] = R$
  - $\Pr[regret \geq 2R] \leq \frac{1}{2}$
    - Markov inequality

MAB: lower bound

- Need to test action a_2
  - Since action a_1 known
- If $\mu_2 = \frac{1}{4}$ and test M times
  - $Regret = \frac{1}{4}M$
  - $\Pr[M \geq 8R] \leq \frac{1}{2}$
- With prob ½ for some 8R seq we stop testing action a_2
  - Assume it is the all zero seq
- Bad event
  - $\mu_2 = \frac{3}{4}$
  - Sequence all zero
  - Probability: $\frac{1}{2} \cdot \left(\frac{1}{4}\right)^{8R}$
  - Regret: $\frac{1}{2} \cdot \left(\frac{1}{4}\right)^{8R} (T - 8R)$
  - Contradiction: for $R = O(\log T)$

Explore-then-Exploit

- Split Explore and Exploit
  - Explore: First kM time steps
    - each action M times
    - kM exploration steps
  - Compute the average of each arm i: $\hat{\mu}_i$
  - Exploit: After time kM
    - Select $\hat{a} = \underset{i}{\operatorname{argmax}}\, \hat{\mu}_i$

Explore-then-Exploit: Algorithm

- Explore
  - For $t \leq kM$
  - $a_t = t \bmod K$

    $T_i = \{t : a_t = i, t \leq KM\}$
    - $\hat{\mu}_i = \frac{1}{M} \sum_{t \in T_i} r_t(i)$
- Exploit:
  - For $t > kM$
    - Use arm $\hat{a} = \underset{i}{\operatorname{argmax}}\, \hat{\mu}_i$

Explore-then-Exploit

- Theorem:

$$E[regret] \leq kM + 2\sqrt{8 \log T / M}\ T + 2/T^2$$

- Corollary:
  - For $M = T^{2/3}$

$$E[regret] \leq kT^{\frac{2}{3}} + 2\sqrt{8 \log T}\ T^{\frac{2}{3}} + 2/T^2 = O(T^{\frac{2}{3}})$$

Explore-then-Exploit: Analysis

- $\mu_i = E[r_t(i)] = E[\hat{\mu}_i]$
- $\Delta_j = \mu^* - \mu_j$
- $E[regret] =$
  - $\sum_{j=1}^{K} \Delta_j M + (T - kM) \sum_{j=1}^{k} \Delta_j \Pr[j = \underset{i}{\operatorname{argmax}}\, \hat{\mu}_i]$
    - Explore: $\sum_{j=1}^{k} \Delta_j M$
    - Exploit: $(T - kM) \sum_{j=1}^{k} \Delta_j \Pr[j = \underset{i}{\operatorname{argmax}}\, \hat{\mu}_i]$

- Concentration bounds
  - Set $\lambda = \sqrt{8 \log T / M}$
  - Then: $\Pr[|\hat{\mu}_i - \mu_i| \geq \lambda] \leq 2 \exp\left(-\frac{\lambda^2 M}{2}\right) = \frac{2}{T^4}$
  - For $k \leq T$: $\Pr[\exists i: |\hat{\mu}_i - \mu_i| \geq \lambda] \leq \frac{2k}{T^4} \leq \frac{2}{T^3}$
  - Bad event: $\text{BAD} = \{\exists i: |\hat{\mu}_i - \mu_i| \geq \lambda\}$
- Assume BAD does not happen
  - The selected action $\hat{a} = i$:
  - $\mu_i + \lambda \geq \hat{\mu}_i \geq \hat{\mu}^* \geq \mu^* - \lambda$
  - $2\lambda \geq \mu^* - \mu_i = \Delta_i$

- Regret
  - Explore: kM
  - Exploit (BAD event does not occur):
    - $(T - kM) \cdot 2\lambda$
  - BAD event:
    - $\Pr[BAD\ event]\, T \leq \frac{2}{T^3} T = \frac{2}{T^2}$
- Total regret:
  - $kM + 2\sqrt{\frac{8 \log T}{M}} T + \frac{2}{T^2}$
- Q.E.D.

Improved regret

- Looking for $\sqrt{T}$ regret
- Explore-Then-Exploit
  - Where did we lose?
  - Sampled very bad and almost optimal the same!
  - Need to have different sample size per arm
    - Depending on its reward

Refined Concentration bounds

- Parameters:
  - $T_t(i) = \{\tau : a_\tau = i, \tau \leq t\}$
  - $n_t(i) = |T_t(i)|$
  - $\hat{\mu}_t(i) = \frac{1}{n_t(i)} \sum_{\tau \in T_t(i)} r_\tau\,(i)$
  - $\lambda_t(i) = \sqrt{\frac{8 \log T}{n_t(i)}}$

- Concentration bound

$$\Pr[|\hat{\mu}_t(i) - \mu_i| \geq \lambda_t(i)] \leq 2\exp\left(-\frac{\lambda_t^2(i)n_t(i)}{2}\right) \leq \frac{2}{T^4}$$

- Good event:
  - $G = \{\forall i\ \forall t\ |\hat{\mu}_t(i) - \mu_i| \leq \lambda_t(i)\}$

- Claim: $\Pr[G] \geq 1 - \frac{2kT}{T^4} \geq 1 - \frac{2}{T^2}$

Confidence Bounds

- Upper Confidence bound
  - $UCB_t(i) = \hat{\mu}_t(i) + \lambda_t(i)$
- Lower confidence bound
  - $LCB_t(i) = \hat{\mu}_t(i) - \lambda_t(i)$
- Assume G occurs
  - $\forall i \; \forall t$: $\mu_i \in [LCB_t(i), UCB_t(i)]$
    - Probability $\geq 1 - \frac{2}{T^2}$

Successive Elimination

- Basic idea
  - Do a round robin over actions
  - If an action has a low average reward
    - Discard it
  - How to set the low average?
    - Compared to the current maximum
  - How to insure that the discards are “safe”
    - With high probability correct

Successive Elimination: Algo

- Maintain a set of actions S
- Initialize $S = K$
- For each phase
  - Use each $i \in S$ once
  - At the end of the phase
    - For each $j \in S$
    - If $\exists i \in S$: $UCB_t(j) < LCB_t(i)$
    - Then $S \leftarrow S - \{j\}$
      - Discard action

Successive Elimination: Analysis

- Desirable properties:
  - Best action never eliminated!
  - Sub-optimal actions eliminated
    - How long sub-optimal action $i \in S$ stays?
      - Bound $n_T(i)$ as a function of Δ_i
- Assume G holds
  - Small additional regret

- Claim: Assuming G holds, the best action never eliminated.
- Proof
  - At any time t
    - $UCB_t(a^*) = \hat{\mu}^* + \lambda_t(a^*) \geq \mu^*$
  - For any sub-optimal action i:
    - $LCB_t(i) = \hat{\mu}_t(i) - \lambda_t(i) \leq \mu_i$
  - Since $\mu^* > \mu_i$
    - $UCB_t(a^*) > LCB_t(i)$
  - Best action never eliminated (under G)
- Q.E.D.

- CLAIM: Assuming G,

suboptimal action i used $n_T(i) \leq \frac{c}{\Delta_i^2} \log T$

- Proof: Assume event G holds:
  - For $i \in S$ at time t:
    - $UCB_t(i) = \hat{\mu}_t(i) + \lambda_t(i) \leq \mu_i + 2\lambda_t(i)$
  - Best action
    - $LCB_t(a^*) = \hat{\mu}^* - \lambda_t(a^*) \geq \mu^* - 2\lambda_t(a^*)$
  - $\lambda_t(i) = \lambda_t(a^*) = \lambda_t$
    - All actions in S selected same number of times
  - $\mu_i + 2\lambda_t \geq \mu^* - 2\lambda_t$

- Let $t = n_T(i)$ be the last time action i used.
- $\lambda_t = \sqrt{8 \log T / n_T(i)}$

❑ Recall

- $\mu_i + 2\lambda_t \geq \mu^* - 2\lambda_t$
  - $4\lambda_t \geq \Delta_i$
  - $4\sqrt{8 \log T / n_T(i)} \geq \Delta_i$
  - $n_T(i) \leq \frac{128}{\Delta_i^2} \log T$

❑ Q.E.D.

Successive Elimination: THM

- Theorem: For the successive elimination algorithm we have
  - $E[regret] \leq \sum_{i \neq a^*} \Delta_i n_T(i) + \frac{2}{T^2} T$
  - $\leq \sum_{i \neq a^*} \frac{c}{\Delta_i} \log T + \frac{2}{T}$
- Corollary:
  - $E[regret] = O(k\sqrt{T} \log T)$
    - For $\Delta_i < 1/\sqrt{T}$, bound $n_T(i) \leq T$

Upper Confidence Bound (UCB)

- *Optimism in face of uncertainty*
- Use the UCB
  - Larger than true expected reward
    - With high probability
  - Errors would decrease the UCB
    - At least the confidence bound λ

UCB: Algorithm

- Initialization: time $t \in [1, k]$
  - At time t let $a_t = t$
    - Each action used once
- For time $t > k$
  - $a_t = \text{argmax}_i \, UCB_{t-1}(i)$

UCB: Analysis

- Assuming the good event G holds
  - $UCB_t(a_t) \geq UCB_t(a^*) \geq \mu^*$
- Suboptimal action i
  - $UCB_t(i) = \hat{\mu}_t(i) + \lambda_t(i) \leq \mu_i + 2\lambda_t(i)$
  - Since $UCB_t(a_t) \geq \mu^*$
    - $\mu^* \leq \mu_i + 2\lambda_t(i)$
    - $\Delta_i \leq 2\sqrt{8 \log T / n_t(i)}$
    - $n_t(i) \leq \left(\frac{32}{\Delta_i^2}\right) \log T$

UCB: Theorem

- Theorem: The UCB algorithm has
  - $E[regret] \leq \sum_{i \neq a^*} \Delta_i E[n_T(i)] + \frac{2}{T^2} T$
  - $\leq \sum_{i \neq a^*} \frac{32}{\Delta_i} \log T + \frac{2}{T}$
- Corollary:
  - $E[regret] = O(k\sqrt{T} \log T)$
    - For $\Delta_i < 1/\sqrt{T}$, bound $n_T(i) \leq T$

UCB versus Successive-Elimination

- Similar amount of exploration
  - Per sub-optimal action
- Difference:
  - When do we explore
- Successive Elimination:
  - Continuous exploration
- UCB
  - Spreading the exploration over time

Best arm identification

- Different goal:
  - Find an optimal action
    - Near optimal
  - Minimize the time of exploration
    - Cost of exploration ignored
      - Only the time

Best arm identification: PAC

- PAC
  - Given $\epsilon, \delta > 0$
  - Return an action i such that
  - With probability $1 - \delta$
    - $\mu^* - \mu_i \leq \epsilon$
  - Goal: Minimize time

Naïve PAC algorithm

- Algorithm:
  - Sample each arm $m(\epsilon, \delta)$ times
    - Error less than ϵ
    - With confidence $1 - \delta/k$
  - Select the action with the best average
    - $\hat{a} = argmax_i \ \hat{\mu}_i$

- How to set $m(\epsilon, \delta)$

Naïve PAC algorithm

- Set $m(\epsilon, \delta) = \frac{8}{\epsilon^2} \log \frac{2k}{\delta}$
  - Concentration bound on averages
    - $\Pr\left[|\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}\right] \leq 2\exp\left(-\frac{\epsilon^2 m}{8}\right) = \frac{\delta}{k}$
  - Union bound over actions
    - $\Pr\left[\exists i: |\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}\right] \leq \delta$
  - Bad event: $\exists i: |\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}$

Naïve PAC algorithm: Analysis

- No Bad event
  - Probability $1 - \delta$
- Bounding
  - $\mu^* - \frac{\epsilon}{2} \leq \hat{\mu}^*$
  - $\hat{\mu}_i \leq \mu_i + \frac{\epsilon}{2}$
  - $\hat{\mu}^* \leq \hat{\mu}_i$

- Assume action i selected
  - $\hat{a} = i$
  - $\hat{\mu}^* \leq \hat{\mu}_i$
- Bound
  - $\mu^* - \frac{\epsilon}{2} \leq \mu_i + \frac{\epsilon}{2}$
  - $\mu^* - \mu_i \leq \epsilon$

Improved Bound

- Naïve PAC algorithm:
  - Total Sample $O(\frac{k}{\epsilon^2}\log\frac{k}{\delta})$
- Successive Elimination
  - Total Sample (worse case): $O(\frac{k}{\epsilon^2}\log\frac{k}{\delta})$
- Median Algorithm
  - Total Sample $O(\frac{k}{\epsilon^2}\log\frac{1}{\delta})$

Median Algo

- Idea
  - Run in phases
  - Each phase sample enough
    - But not enough for the confidence
  - Eliminate the lower half of under-performers
    - Some near-optimal action will remain
      - With high probability
  - Need to re-adjust parameters ϵ, δ

Median Elimination Algorithm

- Input: $\epsilon, \delta > 0$
- Initialize:

$$\ell = 1, S_1 = K, \epsilon_1 = \frac{\epsilon}{4}, \delta_1 = \frac{\delta}{2}$$

- Repeat:
  - For each action $i \in S_\ell$
  - Sample $m(\epsilon_\ell, \delta_\ell)$
    - $\frac{4}{\epsilon_\ell^2} \log \frac{3}{\delta_\ell}$
  - $med = median_{i \in S_\ell}(\hat{\mu}_i)$
  - Update actions
    $S_{\ell+1} = \{i \in S_\ell : \hat{\mu}_i \geq med_\ell\}$
  - Update parameters:
  - $\epsilon_{\ell+1} = \frac{3}{4}\epsilon_\ell$
  - $\delta_{\ell+1} = \frac{\delta_\ell}{2}$
  - $\ell = \ell + 1$
- Until $|S_\ell| = 1$
  - Set $\bar{a} \in S_\ell$

Median Elimination: Analysis

- Parameters:
  - $|S_\ell| = \frac{k}{2^{\ell-1}}$
  - $\epsilon_\ell = \frac{\epsilon}{4}\left(\frac{3}{4}\right)^{\ell-1}$
    - $\sum_\ell \epsilon_\ell \leq \epsilon$
  - $\delta_\ell = \frac{\delta}{2^\ell}$
    - $\sum_\ell \delta_\ell \leq \delta$

- Sample Complexity: $\sum_\ell |S_\ell| \frac{1}{\epsilon_\ell^2} \log \frac{3}{\delta_\ell}$

$$= \sum_\ell \frac{k}{2^{\ell-1}} \frac{64}{\epsilon^2} \log \frac{3 \cdot 2^{\ell-1}}{\delta}$$

$$= O\left(\frac{k}{\epsilon^2} \log \frac{1}{\delta}\right)$$

Median Elimination: Correctness

- Lemma

$$\Pr[\max_{i \in S_\ell} \mu_i \leq \max_{j \in S_{\ell+1}} \mu_j + \epsilon_\ell] \geq 1 - \delta_\ell$$

- Theorem

$$\Pr[\mu^* - \epsilon \leq \mu_{\hat{a}}] \geq 1 - \delta$$

- Proof (of lemma):
  - We do it for $\ell = 1$ (similar for any ℓ)
  - Define bad event $E_1 = \left\{\hat{\mu}^* < \mu^* - \frac{\epsilon_1}{2}\right\}$
    - $\Pr[E_1] \leq \frac{\delta_1}{3}$, assume does not happed
  - Define a bad set of actions
    - $Bad = \{j : \mu^* - \mu_j \geq \epsilon_1, \hat{\mu}_j \geq \hat{\mu}^*\}$
  - For $j \in Bad$
    - $\Pr\left[\hat{\mu}_j \geq \hat{\mu}^* | \hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}\right] \leq \frac{\delta_1}{3}$
      - Conditioning on: $\neg E_1$

- Compute the expected size of Bad
  - $E[|Bad| \,|\neg E_1] \leq k\frac{\delta_1}{3}$
- Bound the probability that Bad is big
  - $\Pr[|Bad| \geq \frac{k}{2}|\neg E_1] \leq \frac{E[|Bad|:\neg E_1]}{\frac{K}{2}} = \frac{2}{3}\delta_1$
- With probability $1 - \delta_1$
  - $\hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}$ and $|Bad| \leq \frac{k}{2}$
- Therefore $\exists j \notin Bad$ and $j \in S_2$
- Q.E.D.

Adversarial MAB

- What happens if there is no distribution on the rewards
  - Adversary sets them!
- We can still use regret
  - How close we are to best action
- Show Vanishing Regret!
- MORE: In the advanced class !!!

MULTI-ARM BANDITS IN PRACTICE

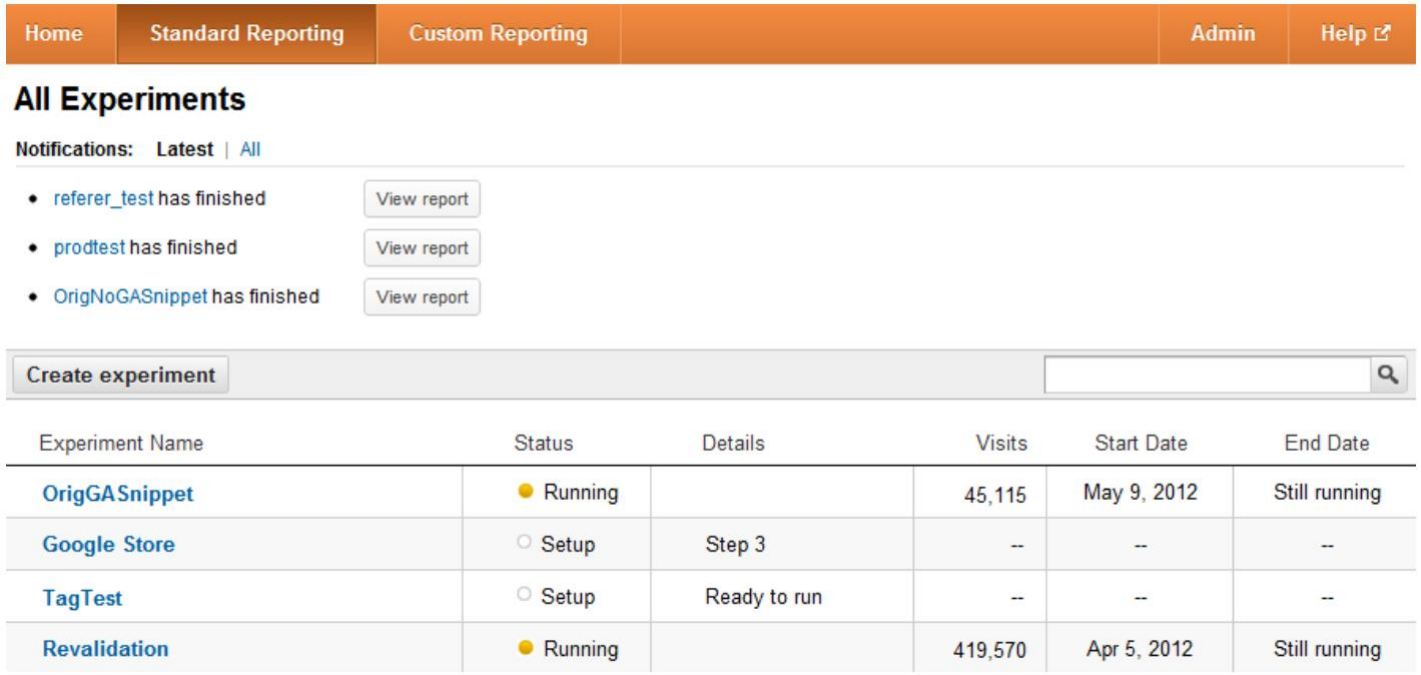
Experiments

- A crucial tool for evaluation
  - Remember: Experiment, Experiment, Experiment
- Widely used in industry
  - Main tool to validate results
  - Experimentation is everywhere in the internet

Content Experiments in Google Analytics

- From Google Analytics website:
  - https://support.google.com/analytics/answer/1745147?hl=en
- With Content Experiments, you can:
  - Compare how different web pages or app screens perform using a random sample of your users
  - Define what percentage of your users are included in the experiment
  - Choose which objective you'd like to test
  - Get updates by email about how your experiment is doing

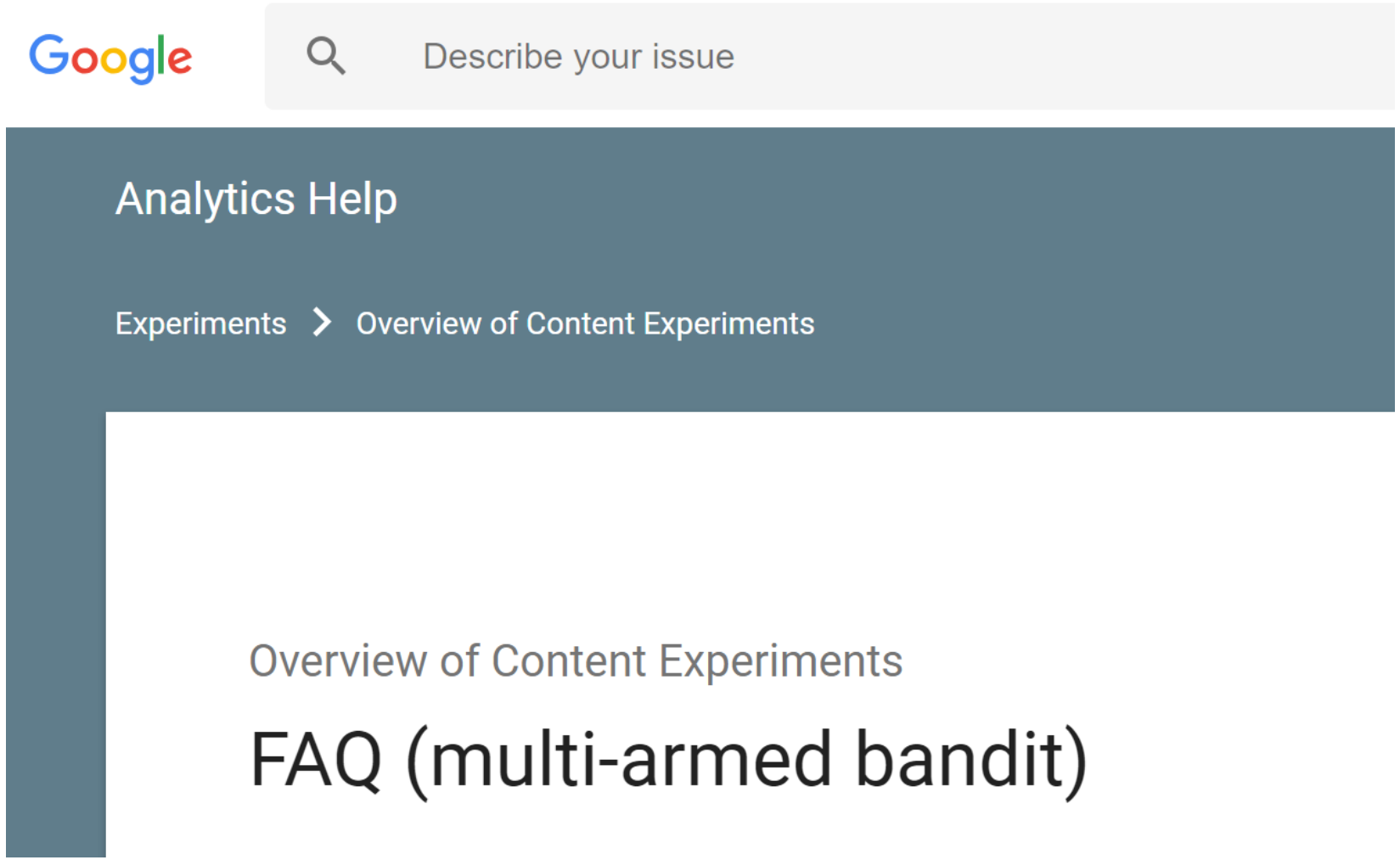
Experiment list



Experiment report

Plot Rows

Variation	Experiment Visits	Conversions	Conversion Rate	Compare to Original	Probability of Outperforming Original
Original	64	34	52.31%	0%	0.00%
Variation 2	70	39	55.71%	7%	66.11%
Variation 3	96	45	46.88%	-10%	25.14%
Variation 1	101	3	2.97%	-94%	0.00%


Google
Describe your issue
Analytics Help
Experiments
Overview of Content Experiments
Overview of Content Experiments
FAQ (multi-armed bandit)

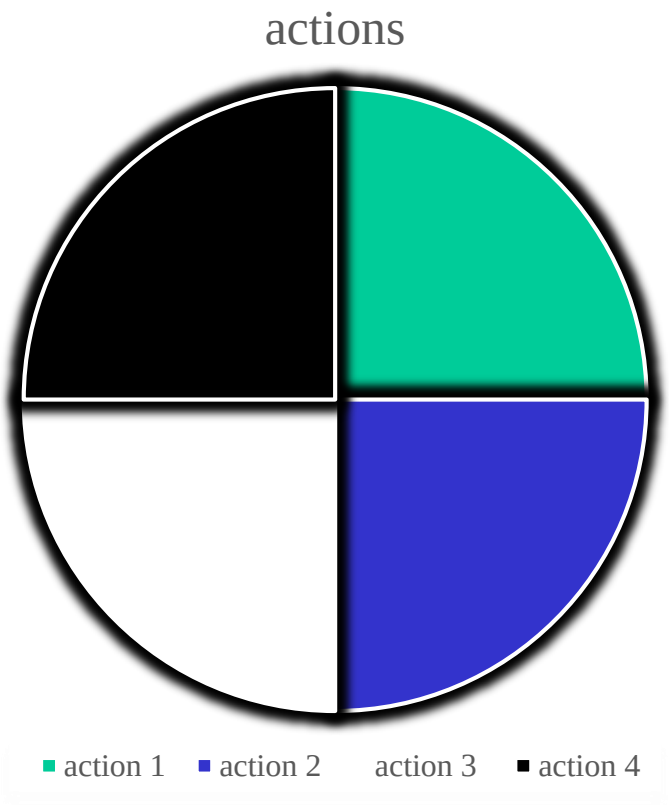
FAQ (multi-armed bandit)

- **Does the bandit always find the optimal arm?**
- **Is it always shorter than a classical test?**
- **What type of experiments make the multi-armed bandit do particularly well (or poorly) compared to classical testing?**
- **What happens if the optimal arm is *unlucky* in the beginning? Can it recover?**
- **Are the results from the bandit statistically valid?**

- https://support.google.com/analytics/answer/2847021?hl=en

A/B Testing versus MAB

A/B testing



Multi-Arm Bandits