Advanced Topics in ML and AGT **Fall Semester, 2018/19**

Lecture 7: December 4, 2018

Lecturer: Yishay Mansour *Scribe[1]: Idan Amir, Guy Tevet, Idan Attias*

1 Stochastic Bandits

[1] In this model, there is a fixed set K of k actions (a.k.a arms). At each round $1 \leq t \leq T$, the player has to choose an action. After choosing it, the player observes the reward of the chosen action, but the rewards of the other actions in K are not revealed to the player.

The reward for action i at round t is denoted by $r_t(i) \sim D_i$, where the reward distribution D_i is over [0,1]. We assume that the rewards are i.i.d. (independent and identically distributed).

Motivation

1. **News**: a user visits a news site and is presented with a news header. The user either clicks on this header or not. The goal of the website is to maximize the number of clicks. So each possible header is an arm in a bandit problem, and the clicks are the rewards
2. **Medical Trials**: Each patient in the trial is prescribed one treatment out of several possible treatments. Each treatment is an arm, and the reward for each patient is the effectiveness of the prescribed treatment.
3. **Ad selection**: In website advertising, a user visits a webpage, and a learning algorithm selects one of many possible ads to display. If ad a is displayed, the website observes whether the user clicks on the ad, in which case the advertiser pays some amount $v_a \in [0, 1]$. So each ad is an arm, and the paid amount is the reward.

Model

- A set of arms $K = \{a_1 \ldots, a_k\}$
- Each arm a_i has a reward distribution D_i
- The expectation of distributaion D_i is:

$$\mu_i = E_{X \sim D_i}[X]$$

- $\mu^* = \max_{i \in K} \mu_i$, and $a^* = \arg\max_{i \in K} \mu_i$
- a_t is the action the player chose at round t

$$Regret = \max_{i \in K} \sum_{t=1}^{T} \underbrace{r_i(t)}_{\text{Random variable}} - \sum_{t=1}^{T} \underbrace{r_{a_t}(t)}_{\text{Random variable}}$$

$$\begin{aligned} \text{Pseudo Regret} &= \max_i E\left[\sum_{t=1}^{T} r_i(t)\right] - E\left[\sum_{t=1}^{T} r_{a_t}(t)\right] \\ &= \mu^* \cdot T - \sum_{t=1}^{T} \mu_{a_t} \end{aligned}$$

[1] Based on the scribes by Emma Rapoport, Shani Einav, Dvir Levi from 2017/2018.

Let T_i be the set of time steps that a_i was chosen:

$$T_i = \{t : a_t = a_i\}$$

The Pseudo Regret can be defined as follows:

$$\text{Pseudo Regret} = \sum_{i \in K} (\mu^* - \mu_i) E[|T_i|]$$

2 Sub-Gaussian Random Variable

For stochastic MAB we will need concentration bounds. For completeness, we will derive the bounds from basics.

Definition 1 *A random variable X is called $\sigma^2 - sub - gaussian$ if for any $\lambda \in \mathbb{R}$:*

$$E\left[e^{\lambda X}\right] \leq e^{\sigma^2 \lambda^2 / 2}$$

Examples

1. $X \sim N\left(0, \sigma^2\right)$; For a normal random variable, $E\left[e^{\lambda X}\right] = e^{\sigma^2 \lambda^2 / 2}$ therefore, X is $\sigma^2 - sub - gaussian$
2. X s.t $\begin{cases} E[X] = 0 \\ |X| \leq B \end{cases}$, then X is $B^2 - sub - gaussian$

Proof: Define $\xi(\lambda) = \log E[e^{\lambda X}]$. We can compute

$$\xi'(\lambda) = \frac{E[X e^{\lambda X}]}{E[e^{\lambda X}]}$$

and

$$\xi''(\lambda) = \frac{E[X^2 e^{\lambda X}]}{E[e^{\lambda X}]} - \left[\frac{E[X e^{\lambda X}]}{E[e^{\lambda X}]}\right]^2$$

We can view $\xi''(\lambda)$ as a variance of the random variable X under the measure $dQ = \frac{e^{\lambda X}}{E[e^{\lambda X}]} dP$, where dP is the original measure of X. Since $|X| \leq B$, then for any distribution $var(X) \leq B^2$. The fundamental theorem of calculus yields,

$$\xi(\lambda) = \int_0^\lambda \int_0^\mu \xi''(\rho) d\rho d\mu \leq B^2 \lambda^2 / 2$$

using $\xi(0) = log(1) = 0$ and $\xi'(0) = E[X] = 0$. □

2.1 Properties of $\sigma^2 - sub - gaussian$ rendom variable X

- $E[X] = 0$, $Var(X) \leq \sigma^2$
- $c \cdot X$ is $c^2\sigma^2 - sub - gaussian$
- If $X_1 \ldots, X_m$ are $\sigma^2 - sub - gaussian$ then $S = \sum_{i=1}^{m} X_i$ is $m\sigma^2 - sub - gaussian$

  Hence, $\frac{1}{m} S = \frac{1}{m} \sum_{i=1}^{m} X_i$ is $\frac{\sigma^2}{m} - sub - gaussian$

Theorem 1 *Let X be $\sigma^2 - sub - gaussian$ random variable, then $Pr[X \geqslant \epsilon] \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$.*

Proof:

$$\begin{aligned} Pr[X \geqslant \epsilon] &= Pr[e^{\lambda X} \geqslant e^{\lambda\epsilon}] \\ &\leq \frac{E[e^{\lambda X}]}{e^{\lambda\epsilon}} \\ &\leq \exp(\sigma^2\lambda^2/2 - \lambda\epsilon), \end{aligned}$$

Where we used Markov's inequality for the first inequality and the fact that the random variable is $\sigma^2 - sub - gaussian$ for the second.

If we choose $\lambda = \frac{\epsilon}{\sigma^2}$, then we get:

$$Pr[X \geqslant \epsilon] \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$$

□

2.2 Hoeffding's inequality

Given $X_1, \ldots, X_m$ i.i.d random variables s.t $X_i \in [0,1]$ and $E[X_i] = \mu$.

Let $\bar{X}_i = X_i - \mu$, then:

$$Pr[\underbrace{\frac{1}{m}\sum_{i=1}^{m} X_i - \mu}_{\frac{1}{m}S} \geq \epsilon] \leq \exp(-\frac{\epsilon^2 m}{2})$$

This is a direct conclusion from last theorem using the facts that $E[\bar{X}_i] = 0$ and:

$$\begin{aligned} X_i \in [0,1] &\Rightarrow |\bar{X_i}| \leq 1 \\ &\Rightarrow X_i \text{ is 1-sub-gaussian} \end{aligned}$$

2.3 Full information $K = 2$

- At time t we observe $\langle r_1(t), r_2(t) \rangle$
- Define
$$S_i(t) = \frac{1}{t}\sum_{\tau=1}^{t} r_i(\tau)$$
- In time $t+1$ we choose:
$$a_{t+1} = \arg\max_{i \in \{1,2\}} S_i(t)$$

Suppose w.l.o.g that $\mu_1 \geq \mu_2$, and let $\Delta = \mu_1 - \mu_2 \geq 0$.

$$\text{Pseudo Regret} = \sum_{t=1}^{\infty} (\mu_1 - \mu_2) Pr\left[S_2(t) \geq S_1(t)\right]$$

$$\forall t : E[S_2(t) - S_1(t)] = \mu_2 - \mu_1 = -\Delta$$

$$Pr\left[\underbrace{S_2(t) - S_1(t) + \Delta}_{\frac{2}{t} - sub - gaussian} \geq \Delta\right] \leq e^{-\Delta^2 \frac{t}{8}}$$

$$\begin{aligned} E\left[\text{Pseudo Regret}\right] &= \sum_{t=1}^{\infty} \Delta Pr\left[S_2(t) \geq S_1(t)\right] \\ &\leq \sum_{t=1}^{\infty} \Delta e^{-\Delta^2 \frac{t}{8}} \\ &\leq \int_0^{\infty} \Delta e^{-\Delta^2 \frac{t}{8}} dt \\ &= \left[\frac{8}{\Delta} e^{-\Delta^2 \frac{t}{8}}\right]_0^{\infty} \\ &= \frac{8}{\Delta} \end{aligned}$$

Notice that this bound doesn't depend on T!

2.4 Bandits

We will now see that we cannot get a regret that does not depend on T for the bandits case. Considering the next example:

$$a_1 \sim Br\left(\frac{1}{2}\right)$$

$$a_2 \sim Br\left(\frac{1}{4}\right)\left(w.p.\frac{1}{2}\right) \qquad or \qquad a_2 \sim Br\left(\frac{3}{4}\right)\left(w.p.\frac{1}{2}\right)$$

Assume by way of contradiction

$$E\left[\sum_{i\in\{1,2\}} \Delta_i |T_i|\right] = E\left[PseudoRegret\right] = R$$

where R does not depend on T.
By Markov inequality:

$$Pr\left[PseudoRegret \geq 2R\right] \leq \frac{1}{2}$$

Since μ_1 is known, an optimal algorithm will first check a_2 in order to decide which action is better and stick with it.

Assuming $\mu_2 = \frac{1}{4}$, and the algorithm decided to stop playing a_2 after M rounds, Then:

$$PseudoRegret = \frac{1}{4}M$$

Thus,

$$Pr\left[PseudoRegret \geq 2R\right] = Pr\left[M \geq 8R\right] \leq \frac{1}{2}$$

And,

$$Pr\left[M < 8R\right] > \frac{1}{2}$$

Hence, the probability that after $8R$ rounds, the algorithm will stop playing a_2 (if $\mu_2 = \frac{1}{4}$) is at least $\frac{1}{2}$.

Assuming $\mu_2 = \frac{3}{4}$, but all $8R$ first rounds, playing a_2 yield the value zero (with probability $\left(\frac{1}{4}\right)^{8R}$), the algorithm will most probably stop playing a_2, even though it is the preferred action. In this case, we will get:

$$PseudoRegret = \frac{1}{4}(T - M) \approx \frac{1}{4}T$$

The expected Pseudo Regret is,

$$E\left[PseudoRegret\right] = R \geq \underbrace{\frac{1}{2}}_{a_2 \sim Pr(Br(\frac{3}{4}))} \cdot \underbrace{\left(\frac{1}{4}\right)^{8R}}_{Pr(all\ 0 | a_2 \sim Pr(Br(\frac{3}{4})))} \cdot (T - 8R) \approx e^{-R}T$$

Which implies that:

$$R = O\left(\log T\right)$$

Contrary to the assumption that R does not depend on T. [2]

3 Explore-Then-Exploit

- We choose a time frame kM to explore.
- We explore each arm M times - kM rounds of exploration.
- After kM rounds we choose the action with the highest average reward during the explore phase.

Define:

$$T_j = \{t : t \mod K = j, t \leq K \cdot M\}$$

$$\hat{\mu}_j = \frac{1}{M}\sum_{t \in T_j} r_j(t)$$

$$\mu_j = E\left[r_j(t)\right]$$

Then we define the distance of each action from the optimal action:

$$\Delta_j = \mu^* - \mu_j$$

$$E\left[\text{Pseudo regret}\right] = \underbrace{\sum_{j=1}^{K} \Delta_j \cdot M}_{Explore} + \underbrace{(T - K \cdot M)\sum_{j=1}^{K} \Delta_j Pr\left[\hat{\mu}_j = \max_i \hat{\mu}_i\right]}_{Exploit}$$

let,

$$\lambda = \sqrt{\frac{8\log T}{M}}$$

Therefore,

$$Pr\left[|\hat{\mu}_j - \mu_j| \geq \lambda\right] \quad \leq \quad 2e^{-\frac{\lambda M}{2}} = \frac{2}{T^4}$$

Using a union bound,

$$Pr\underbrace{\left[\exists_j : |\hat{\mu}_j - \mu_j| \geq \lambda\right]}_{B} \quad \leq \quad \frac{2k}{T^4} \underset{for\ k \leq T}{\leq} \frac{2}{T^3}$$

[2] More formally, after $8R$ steps, there exists some sequence of outcomes that cause the algorithm to switch to action a_2. The probability of that sequence is at least $\left(\frac{1}{4}\right)^{8R}$.

Let the "bad event" $B = \exists_j : |\hat{\mu}_j - \mu_j| \geq \lambda$. If B didn't happen, then for each j:

$$\mu_j + \lambda \geq \hat{\mu}_j \geq \hat{\mu}^* \geq \mu^* - \lambda$$

therefore:

$$2\lambda \geq \mu^* - \mu_j = \Delta_j$$

and therefore:

$$\Delta_j \leq 2\lambda$$

Then, we get the following regret:

$$\begin{aligned} E\left[\text{Pseudo regret}\right] &\leq \underbrace{\left(\sum_{j=1}^{K} \Delta_j\right) M}_{Explore} + \underbrace{(T - K \cdot M) \cdot 2\lambda}_{\text{B didn't happen}} + \underbrace{\frac{2}{T^3} \cdot T}_{\text{B happened}} \\ &\leq K \cdot M + 2 \cdot \sqrt{\tfrac{8 \log T}{M}} \cdot T + \tfrac{2}{T^2} \end{aligned}$$

If we optimize the number of exploration phases M and choose $M = T^{\frac{2}{3}} K^{-\frac{2}{3}}$, we get:

$$K^{\frac{1}{3}} \cdot T^{\frac{2}{3}} + 2 \cdot \sqrt{8 \log T} \cdot K^{\frac{1}{3}} T^{\frac{2}{3}} + \frac{2}{T^2}$$

which is sublinear but more than $O\left(\sqrt{T}\right)$.

4 More Advanced Algorithms

- We will look at some more advanced algorithms:

Define:
$n_i(t)$ - the number of times we chose action i by round t
$\hat{\mu}_i(t)$ - the average reward of action so far, that is:

$$\hat{\mu}_i(t) = \frac{1}{n_i(t)} \sum_{\tau=1}^{t} r_i(\tau) \mathbb{I}\left(a_\tau = i\right)$$

We observe that $n_i(t)$ is a random variable and not a number!
We would like to get the following result:

$$Pr\left[|\hat{\mu}_i(t) - \mu_i| \geq \underbrace{\sqrt{\frac{8 \log T}{n_i(t)}}}_{\lambda_i(t)}\right] \leq \frac{2}{T^4}$$

Because $n_i(t)$ is not a number, we cannot use a concentration bound right away. Instead, we would like to use concentration bound for each possible average at each possible time (even if it did not actually happen).

We look at the m^{th} time we sampled action i:

$$\hat{\mathbb{V}}_i(m) = \frac{1}{m} \sum_{\tau=1}^{m} r_i(t_\tau^i)$$

Where the t_τ^i's is the round when we choose action i for the τ time.

Now we fix m and get:

$$\forall_i \forall_m Pr\left[\left|\hat{\mathbb{V}}_i(m) - \mu_i\right| \geq \sqrt{\frac{8\log T}{m}}\right] \leq \frac{2}{T^4}$$

and notice that $\hat{\mu}_i(t) \equiv \hat{\mathbb{V}}_i(m)$ when $m = n_i(t)$.
Define the "good event" G:

$$G = \{\forall_i \forall_t \, |\hat{\mu}_i(t) - \mu_i| \leq \lambda_i(t)\}$$

From the concentration bound,

$$Pr\,(G) \geq 1 - \frac{2}{T^2}$$

5 Confidence Bound

Define the upper confidence bound:

$$UCB_i(t) = \hat{\mu}_i(t) + \lambda_i(t)$$

and similarly, the lower confidence bound:

$$LCB_i(t) = \hat{\mu}_i(t) - \lambda_i(t)$$

Assuming G holds then:

$$\forall_i \forall_t \mu_i \in [LCB_i(t), UCB_i(t)]$$

Therefore:

$$Pr\left[\forall_i \forall_t \mu_i \in [LCB_i(t), UCB_i(t)]\right] \geq 1 - \frac{2}{T^2}$$

5.1 Successive Elimination

We maintain a set of actions S.
Initially $S = K$
In each phase:

- We try every $i \in S$ once
- For each $j \in S$ if there exists $i \in S$ such that:

$$UCB_i(t) < LCB_j(t)$$

We remove j from S, that is we update:

$$S \leftarrow S - \{j\}$$

We will get the following results:

- As long as action i is still in S, we've tried it exactly the same number of times as all of the other actions still in S.
- The best action [under the assumption of G] is never eliminated from S.

Under the assumption of G we get:

$$\mu^* - 2\lambda^\star \leq \hat{\mu}^* - \lambda^\star = LCB_* < UCB_i = \hat{\mu_i} + \lambda_i \leq \mu_i + 2\lambda_i$$

where $\lambda = \lambda_i = \lambda^*$ because we've chosen action i and the best action the same number of times so far.

Therefore:

$$\Delta_i = \mu^* - \mu_i \leq 4\lambda = 4\sqrt{\frac{8\log T}{n_i(t)}}$$

$$\Rightarrow \; n_i(T) \leq \frac{c}{\Delta_i^2}\log T$$

$$\Rightarrow \; E\left[\text{Pseudo Regret}\right] \;=\; \sum_{i=1}^{k} \Delta_i n_i(T)$$

$$\leq \; \sum_{i=1}^{k} \frac{c}{\Delta_i}\log T + \underbrace{\frac{2}{T^2}\cdot T}_{\text{The bad event}}$$

meaning that the expected pseudo regret is bounded by $O\left(\sum_{i=1}^{k} \frac{c}{\Delta_i}\log T\right)$.

5.2 Upper confidence bound (UCB)

- We try each action once (for a total of k rounds)
- Afterwards we choose:

$$a_t = \arg\max_{i \in K} UCB_i$$

If we chose action i then necessarily [under the assumption of G]:

$$UCB_i \geq UCB_* \geq \mu^*$$

For any action i:

$$UCB_i = \hat{\mu}_i + \lambda_i \leq \mu_i + 2\lambda_i$$

Combining the two inequalities we get,

$$\mu_i + 2\lambda_i \geq \mu^*$$

$$\Rightarrow \; 2\lambda_i \geq \mu^* - \mu_i = \Delta_i$$

Then each time we've chosen action i, we couldn't have made a very big mistake because:

$$\Delta_i \leq \; 2 \cdot \sqrt{\frac{8\log T}{n_i(t)}}$$

And therefore if i is very far off from the optimal action, we wouldn't choose it too many times, because:

$$n_i(t) \leq \frac{c}{\Delta_i^2}\log T$$

And over all we get:

$$E\left[\text{Pseudo Regret}\right] = \sum_{i=1}^{k} \Delta_i E\left[n_i(T)\right] + \underbrace{\frac{2}{T^2} \cdot T}_{\text{The bad event}}$$

$$\leq \sum_{i=1}^{k} \frac{c}{\Delta_i} \cdot \log T + \frac{2}{T}$$

Note the similarity between the regret bounds of Successive Elimination and Upper Confidence Bound.

6 Best Arm Identification

First Goal Given $\epsilon, \delta > 0$, find in probability $1-\delta$, an action i such that:

$$\mu^* - \mu_i \leq \epsilon$$

Second Goal Given $\Delta \leq \mu^* - \mu_i$ (for every i that isn't optimal), find the optimal action a_* in probability $1-\delta$.

The first goal is similar in spirit to PAC learning, and we will focus on it.

6.1 Naive Algorithm (for the first goal):

We sample each action i $m = O\left(\frac{1}{\epsilon^2}\log\frac{k}{\delta}\right)$ times, and return $\hat{a^\star} = \arg\max_i \hat{\mu}_i$.

If all the rewards are between 0 to 1, then for every action i:

$$Pr\left[\underbrace{|\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}}_{\text{bad event}}\right] \leq e^{-\left(\frac{\epsilon}{2}\right)^2 m/2} = \frac{\delta}{k}$$

by union bound we get:

$$Pr\left[\exists_i \, |\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}\right] \leq \delta$$

If the bad event didn't happen, then:

$$\begin{cases} \mu^* - \frac{\epsilon}{2} \leq \hat{\mu}^* \\ \mu_i + \frac{\epsilon}{2} \leq \hat{\mu}_i \end{cases}$$

If the returned action is $i = a = \arg\max_j \hat{\mu}_j$, then:

$$\Rightarrow \ \mu_i + \frac{\epsilon}{2} \geq \hat{\mu}_i \geq \hat{\mu}^* \geq \mu^* - \frac{\epsilon}{2}$$

$$\Rightarrow \ \epsilon \geq \mu^* - \mu_i$$

And therefore $a = \arg\max_i \hat{\mu}_i$ is the optimal action in probability $1-\delta$.

We would like to improve this algorithm by reducing the number of trials from $O\left(\frac{k}{\varepsilon^2}\log\frac{k}{\delta}\right)$ to $O\left(\frac{k}{\varepsilon^2}\log\frac{1}{\delta}\right)$.

6.2 Median Algorithm:

The idea: the algorithm runs for l phases, after each phase we eliminate half of the actions. This elimination allows us to sample each action more times in the next phase which makes eliminating the near optimal action unlikely.

Algorithm 1 Best Arm Identification

Input: $\epsilon, \delta > 0$
Output: $\bar{a} \in K$
Init: $S_1 = K$, $\epsilon_1 = \frac{\varepsilon}{4}$, $\delta_1 = \frac{\delta}{2}$, $l = 1$
Repeat:
$\forall_i \in S_l$, sample action i, $m_l = \frac{1}{\left(\frac{\epsilon_l}{2}\right)^2} \log\left(\frac{3}{\delta_l}\right)$ times
$\hat{\mu}_i \leftarrow$ mean (only of samples during the l^{th}phase)
$\text{median}_l \leftarrow median\left\{\hat{\mu}_i : i \in S_l\right\}$
$S_{l+1} \leftarrow \left\{i \in S_l : \hat{\mu}_i \geq \text{median}_l\right\}$
$\epsilon_{l+1} \leftarrow \frac{3}{4}\epsilon_l$
$\delta_{l+1} \leftarrow \frac{\delta_l}{2}$
$l \leftarrow l + 1$
Until $|S_l| = 1$

Complexity: During phase l we have $|S_l| = \frac{k}{2^{l-1}}$ actions.

$$\epsilon_l = \frac{\epsilon}{4}\left(\frac{3}{4}\right)^{l-1}, \ \delta_l = \frac{\delta}{2^l}$$

Therefore,

$$\sum \epsilon_l \leq \epsilon, \ \sum \delta_l \leq \delta$$

The total number of rounds is therefore:

$$\begin{aligned}\sum_l |S_l| \cdot m_l = \sum_l \frac{k}{2^{l-1}} \frac{64}{\epsilon^2}\left(\frac{16}{9}\right)^{l-1} \log\frac{3 \cdot 2^l}{\delta} &= \sum_l k\left(\frac{8}{9}\right)^{l-1}\left[c \cdot \frac{\log\frac{1}{\delta}}{\epsilon^2} + \frac{\log 3}{\epsilon^2} + \frac{l}{\epsilon^2}\right] \\ &= O\left(\tfrac{k}{\epsilon^2}\log\tfrac{1}{\delta}\right)\end{aligned}$$

Theorem 2 $Pr\left[\underbrace{\max_{j \in S_l} \mu_j}_{action\ l} \leq \underbrace{\max_{j \in S_{l+2}} \mu_j}_{action\ l+1} + \epsilon_l\right] \geq 1 - \delta_l$

Proof: We will prove for $l = 1$, but similarly this holds for any l.

Define the event: $E_1 = \left\{\hat{\mu}^* < \mu^* - \frac{\epsilon_1}{2}\right\}$. Then,

$$Pr\left[E_1\right] \leq e^{-(\frac{\epsilon_1}{2})^2 m/2} \leq \frac{\delta_1}{3}$$

If E_1 didn't happen, we define a bad set:

$$\text{Bad} = \left\{j \ : \ \mu^* - \mu_j \geq \epsilon_1, \ \hat{\mu}_j \geq \hat{\mu}^*\right\}$$

Consider an action j such that $\mu^\star - \mu_j \geq \epsilon_1$. Then:

$$Pr[\hat{\mu}_j \geq \hat{\mu}^* \ \underbrace{\left[\hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}\right]}_{\neg E_1}] \leq Pr\left[\hat{\mu}_j \geq \mu_j + \frac{\epsilon_1}{2} \,|\neg E_1\right] \leq \frac{\delta_1}{3}$$

We now compute the expected size of Bad, given that E_1 does not hold.

$$E\left[|\text{Bad}| \ \ |\neg E_1\right] \leq k\frac{\delta_1}{3}$$

Using Markov's inequality we get:

$$\begin{aligned} Pr\left[|\text{Bad}| \geq \tfrac{k}{2} | \neg E_1\right] &\leq \tfrac{E[|\text{Bad}||\neg E_1]}{k/2} \\ &= \tfrac{2}{3}\delta_1 \end{aligned}$$

This implies that with probability $1-\delta_1$: $\hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}$ and $|\text{Bad}| \leq \frac{k}{2}$.

Therefore, there exist an action $j \notin \text{Bad}$ and $j \in S_2$, which proves the theorem. □