Ans 33.1 $R_n^{\text{Sample}}(\pi, \nu) = \frac{R_n(\pi, \nu)}{n}$

where $R_n(\pi, \nu)$ is the cumulative regret of policy π on environment ν.

Proof :-

Let $(\pi)_{t=1}^{n}$ be a policy & let

$$\pi(i \mid a_{1,1}, \ldots a_{n,n}) = \frac{1}{n} \sum_{t=1}^{n} 1\{a_t = i\}$$

Thus by regret decomposition

$$R_n(\pi, \nu) = E\left[\sum_{i=1}^{n} \Delta_i N_i(n)\right]$$

$$= n\, E\left[\sum \Delta_i \frac{N_i(n)}{n}\right]$$

$$= n\, E\left[\Delta_{A_{n+1}}\right]$$

$$= n\, R_n^{\text{Sample}}\left((\pi_t)_1^{n+1}, \nu\right)$$

$$\Rightarrow R_n^{\text{Sample}}(\pi, \nu) = \frac{R_n(\pi, \nu)}{n}$$

We know that minimax bound on R_n is

$$R_n(\pi, \nu) \geq c\sqrt{(K-1)n} \geq c\sqrt{Kn}$$

$$R_n^{\text{Sample}}(\pi, \nu) \geq c\sqrt{\frac{K}{n}}$$

33.2. As proved above,

$$R_n^{\text{sample}}(\pi, \nu) = \frac{R_n(\pi, \nu)}{n}$$

For a UE policy,

For any bandit, no of optimal pulls $\geq (K-1)\left\lfloor \frac{n}{K} \right\rfloor$

$$R_n(\pi, \nu) \geq (K-1)\left\lfloor \frac{n}{K} \right\rfloor \sum_{\substack{i=1 \\ i \neq i^*}}^{K} \Delta_i$$

Assuming 1 subgaussian rewards,

The term on RHS $\geq c\sqrt{nK\log K}$

$$R_n^{\text{sample}} \geq c\sqrt{\frac{K\log K}{n}}$$