CS246 Exam 2025 — Question 12: Bandits (10 points)

Part 1 (2 points) — Regret Is Unknown While Running?

Answer: (a) True

Regret is $R_T = \Sigma(\mu^* - \mu_{\{i_t\}})$, where μ^* is the mean reward of the **best arm**. But μ^* is unknown — if we knew which arm was best, there'd be no bandit problem. Regret can only be computed retrospectively (e.g., in simulations where true parameters are known to the experimenter but hidden from the algorithm).

Part 2 (2 points) — UCB Won't Pick Suboptimal Arms?

Answer: (b) False

UCB selects by: $UCB(a) = \hat{\mu}_a + \alpha\sqrt{(2 \ln t / m_a)}$. If an arm hasn't been played for a while, its m_a stays fixed while t grows, so its confidence bonus increases. Eventually a neglected arm's UCB exceeds the currently exploited arm's UCB, and it gets selected again. **No arm is ever permanently ruled out** — this is by design ("optimism in the face of uncertainty") and is what ensures UCB achieves sublinear regret.

Part 3 (2 points) — Standard Statistical Tests Apply to Bandits?

Answer: (b) False

Standard tests (t-tests, etc.) assume **fixed, non-adaptive** random sampling — each subject is independently assigned to a group with pre-determined allocation. In bandit experiments, the allocation is **adaptive**: which arm gets traffic depends on past outcomes. This violates the independence assumptions, producing miscalibrated p-values. Bandit algorithms instead rely on their own regret bounds for convergence guarantees.

Part 4 (4 points) — ε-Greedy Computation

Data from 10 rounds:

Round	Arm	Reward
1	1	1
2	2	0
3	3	1
4	1	0
5	2	1
6	3	0
7	1	1
8	2	1
9	3	0
10	2	1

Part 4a — Mean Rewards

Group rewards by arm and compute empirical means:

Arm	Rewards	Mean
1	1, 0, 1 (3 pulls)	**2/3 ≈ 0.667**
2	0, 1, 1, 1 (4 pulls)	**3/4 = 0.75**
3	1, 0, 0 (3 pulls)	**1/3 ≈ 0.333**

Part 4b — Probabilities for Round 11

The best arm is **Arm 2** (highest mean = 3/4). With $\varepsilon = 0.3$ and 3 arms:

- **Exploit** (probability $1 - \varepsilon = 0.7$): choose best arm (Arm 2) deterministically
- **Explore** (probability $\varepsilon = 0.3$): choose uniformly at random → each arm gets $0.3/3 = 0.1$

Combining:

Arm	Exploit	+ Explore	= Total
1	0	+ 0.1	**0.1**
2	0.7	+ 0.1	**0.8**
3	0	+ 0.1	**0.1**

Verification: 0.1 + 0.8 + 0.1 = 1.0 ✓