1. Stochastic bandits.
Stochastic bandit problems.

- $k$ arms.
- Some model for reward distributions $P_\theta$ for $\theta \in \Theta$. (But $\Theta$ might be very large; e.g., \{ $P_\theta : \theta \in \Theta$ \} might be the set of all probability distributions on $[0, 1]$.)
- Arm $j$ has unknown reward distribution $P_{\theta_j}$, and pulling that arm produces rewards $X_{j,1}, X_{j,2}, \ldots$ chosen independently from $P_{\theta_j}$.
- At time $t$, the problem is to use the available information (that is, previous choices and outcomes, $I_1, X_{I_1,1}, \ldots, I_{t-1}, X_{I_{t-1},t-1}$) to choose an arm $I_t \in \{1, \ldots, k\}$.
- This choice can be randomized.
Stochastic bandit problems.

We aim to get a high total reward. Several formulations:

1. We might consider regret,

\[ R_n = \max_{j^* = 1, \ldots, k} \sum_{t=1}^{n} X_{j^*, t} - \sum_{t=1}^{n} X_{I_t, t}, \]

and aim to minimize expected regret, \( \mathbb{E} R_n \), or aim to minimize regret with high probability,

\[ \Pr(R_n - f_n \geq \epsilon) \leq \delta. \]
2. Or we might consider total reward,

\[ \sum_{t=1}^{n} X_{I_t,t}. \]

Maximizing expected total reward is equivalent to minimizing pseudo-regret,

\[
\overline{R}_n = \max_{j^* = 1, \ldots, k} \mathbb{E} \left[ \sum_{t=1}^{n} X_{j^*,t} - \sum_{t=1}^{n} X_{I_t,t} \right]
\]

\[ = n \max_{j^* = 1, \ldots, k} \mu_{j^*} - \mathbb{E} \sum_{t=1}^{n} X_{I_t,t}, \]

where \( \mu_{j} = \mathbb{E} X_{j,1}. \) Note that \( \overline{R}_n \leq \mathbb{E} R_n. \) We might instead aim to maximize total reward with high probability.
Fluctuations in $\sum_{t=1}^{n} X_{j,t}$ grow like $\sqrt{n}$, so we cannot hope to achieve $\mathbb{E}R_n$ better than this order. We’ll focus on pseudo-regret.

Notation:

- Mean reward: $\mu_j = \mathbb{E}X_{j,1}$.
- Best: $\mu^* = \max_{j^* = 1, \ldots, k} \mu_{j^*}$.
- Gap: $\Delta_j = \mu^* - \mu_j$.
- Number of plays: $T_j(s) = \sum_{t=1}^{s} 1[I_t = j]$.

Hence,

$$\overline{R}_n = n\mu^* - \sum_{j=1}^{k} \mathbb{E}T_j(n)\mu_j = \sum_{j=1}^{k} \mathbb{E}T_j(n)\Delta_j.$$