18-661: Introduction to ML for Engineers

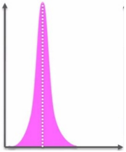
Multi-Armed Bandits

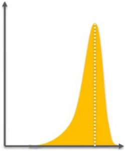
Spring 2025

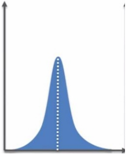
ECE – Carnegie Mellon University

Multi-Armed Bandit

\$1	\$1	\$1
\$0	\$4	\$1
\$0	\$0	\$0
\$1	\$2	\$2







Stochastic Bandit: Setting

At each time-step $t = 1, 2, \ldots, T$,

Arm 1

Arm 2

Arm 3

Stochastic Bandit: Setting

At each time-step $t = 1, 2, \ldots, T$,

pull/play arm $i_t \in \{1, \ldots, n\}$

Arm 1

Arm 2

Arm 3

Stochastic Bandit: Setting

At each time-step $t = 1, 2, \ldots, T$,

pull/play arm $i_t \in \{1, \ldots, n\}$

and receive reward $r(i_t)$

Arm 1

Arm 2

Arm 3

Stochastic Bandit: Performance Metric

Arm 1

Arm 2

Arm 3

Goal: maximize total reward accumulated over time

Stochastic Bandit: Performance Metric

Arm 1

Arm 2

Arm 3

Goal: maximize total reward accumulated over time

Performance Metric: Regret

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} r(i^\star)\right] - \mathbb{E}\left[\sum_{t=1}^{T} r(i_t)\right]$$

Stochastic Bandit: Performance Metric

Arm 1

Arm 2

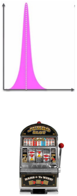
Arm 3

Goal: maximize total reward accumulated over time

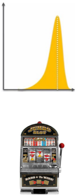
Performance Metric: Regret

$$
\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T} r(i^\star)\right] - \mathbb{E}\left[\sum_{t=1}^{T} r(i_t)\right] \\
&= T\rho^\star - \mathbb{E}\left[\sum_{t=1}^{T} r(i_t)\right]
\end{aligned}
$$

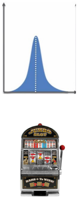
Exploration - Exploitation Tradeoff



$1

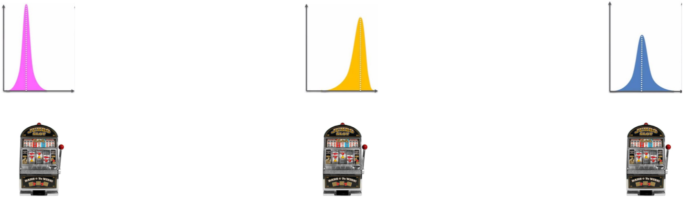


-



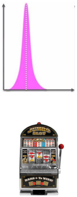
-

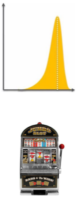
Exploration - Exploitation Tradeoff

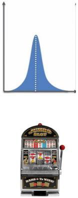


$1	-	-
-	$1	-

Exploration - Exploitation Tradeoff

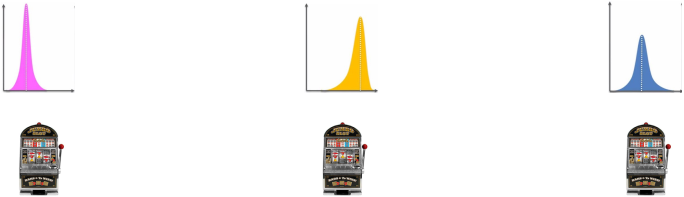






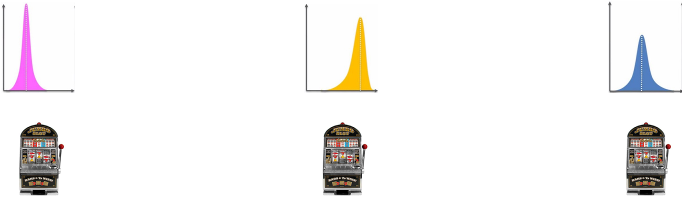
$1	-	-
-	$1	-
-	-	$1

Exploration - Exploitation Tradeoff



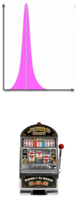
\$1	-	-
-	\$1	-
-	-	\$1
\$0	-	-

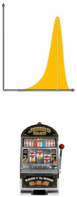
Exploration - Exploitation Tradeoff

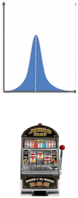


$1	-	-
-	$1	-
-	-	$1
$0	-	-
-	-	$1

Exploration - Exploitation Tradeoff

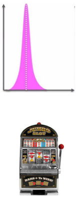


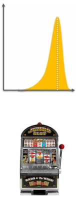


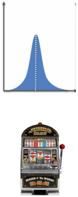


$1	-	-
-	$1	-
-	-	$1
$0	-	-
-	-	$1
-	-	$0

Exploration - Exploitation Tradeoff

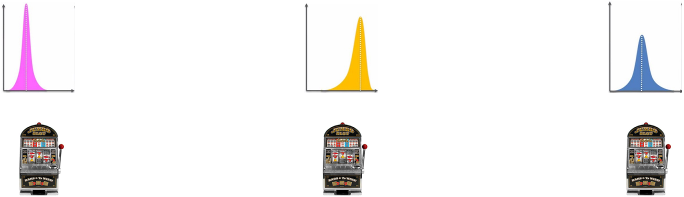






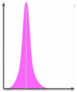
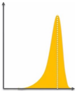
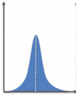
$1	-	-
-	$1	-
-	-	$1
$0	-	-
-	-	$1
-	-	$0
-	$4	-

Exploration - Exploitation Tradeoff



$1	-	-
-	$1	-
-	-	$1
$0	-	-
-	-	$1
-	-	$0
-	$4	-
-	$0	-

Exploration - Exploitation Tradeoff

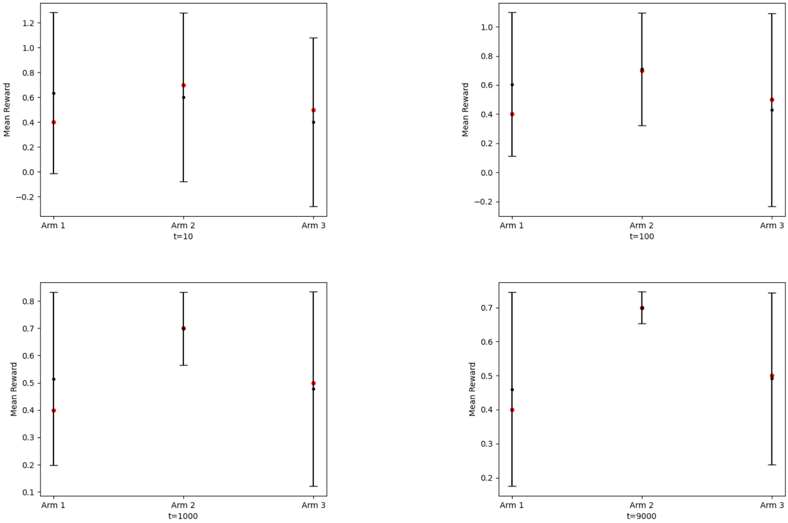
$1	-	-
-	$1	-
-	-	$1
$0	-	-
-	-	$1
-	-	$0
-	$4	-
-	$0	-
-	$2	-

UCB: Optimism in the face of Uncertainty

Algorithm 1 UCB

```
for $t = 1, 2, \ldots, T$ do
    Play arm $i_t = \arg\max_i \mathrm{UCB}_{i,t} = \left( \frac{\sum_{u=0}^{t} r(i_u) \mathbb{1}_{i_u = i}}{T_i} + \sqrt{\frac{2 \log t}{T_i}} \right)$
    Observe reward $r_{i_t}$
end for
```

UCB: Optimism in the face of Uncertainty



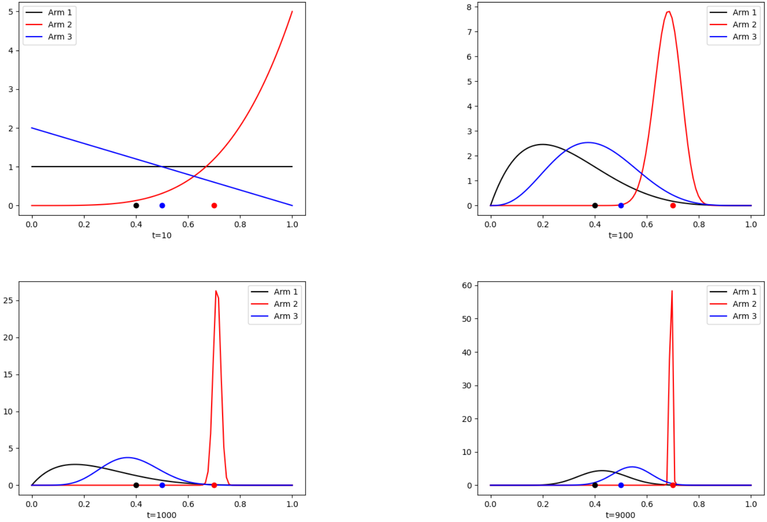
Bernoulli Bandit with means $0.4, 0.7, 0.5$

Thompson Sampling

Algorithm 2 Thompson Sampling

```
for $t = 1, 2, \dots, T$ do
    Sample $\hat{\mu}_{i,t} \sim P_{i,t-1}$ for each arm $i \in \{1, \dots, n\}$
    Play arm $i_t = \arg\max_i \hat{\mu}_{i,t}$
    Observe reward $r_{i_t,t}$ and update posterior $P_{i,t}$
end for
```

Thompson Sampling



Bernoulli Bandit with means $0.4, 0.7, 0.5$