Bandits

Multi-armed adversarial bandits, stochastic bandits, contextual bandits

Chu Duc Thang - October 9th

With many slides derived from Haipeng Lou, David Silver, Emma Brunskill, and

Outline

1. **Adversarial bandit**
2. Stochastic bandit
3. Bayesian bandit
4. Contextual bandit

1. Adversarial bandit

Setting - Notation

$w_t \in \Delta(K)$ - Action in "Expert advice" problem

$f_t(a_t) = \langle l_t, a_t \rangle$ - Loss in "Expert advice" problem

- **Action at time t:** $\boxed{a_t \in [K]}$ ← Special case of $w_t = [0,\ldots,1,\ldots,0]$
- **Loss at time t:** $l_t \in [0,1]^K$
  - **Remark**: Only know $\boxed{l_t(a_t)}$ ← Partial information
- **Setting**: Adversarial

1. Adversarial bandit

Setting - Regret

$R_T = \sum_{t=1}^{T} \langle w_t, l_t \rangle - \min_{w \in \Omega} \langle w_t, l_t \rangle$ - Regret in "Expert advice" problem

- **Regret:** $\boxed{\mathbb{E}[R_T]} = \mathbb{E}[\sum_{t=1}^{T} l_t(a_t)] - \min_{a \in [K]} \sum_{t=1}^{T} l_t(a)$
  - **Remark:** The expectation due to randomness of the algorithm

1. Adversarial bandit

Setting - Exploration vs Exploitation

- **Applications**: Clinical trials, recommendation systems, etc.
- **Foundation model of Exploration-Exploitation**:
    - **Exploitation**: Select a_t to minimize losses in the past
    - **Exploration**: Select a_t leading to even smaller losses

1. Adversarial bandit
Algorithm - Motivation

- **Question**: Since problem setting is **very similar** to Expert problem setting, can we use Hedge algorithm to solve?
- **Answer**: Yes (with modifications)

1. Adversarial bandit
Algorithm - Exp3

Algorithm 2: Hedge

Input: learning rate $\eta > 0$ (also called step size, temperature, etc.)
Initialization: let $L_0 \in \mathbb{R}^N$ be the all-zero vector
for $t = 1, \ldots, T$ **do**
  compute $p_t \in \Delta(N)$ such that $p_t(i) \propto \exp(-\eta L_{t-1}(i))$
  play p_t and observe loss vector $\ell_t \in [0,1]^N$
  update $L_t = L_{t-1} + \ell_t$

- **Estimator** $\hat{l}_t \in \mathbb{R}_+^K$:
  - **Purpose**: Only know $l_t(a_t)$, while Hedge requires full l_t
  - **Inverse importance weighted estimator**: Pick $a_t \sim p_t \in \Delta(K)$, $\hat{l}_t(a) = \dfrac{l_t(a)}{p_t(a)} 1\{a = a_t\}$
  - **Algorithm:**
    - Initialize $L_0 = (0,\ldots,0)$
    - For $t = 1,\ldots,T$
      - Compute $p_t \in \Delta(K)$ such that $p_t(a) \propto \exp(-\eta \sum_{s<t} \hat{l}_s(a)), \forall a \in [K]$
      - Play $a_t \sim p_t \in \Delta(K)$ and observe $l_t(a_t)$
      - Calculate $\hat{l}_t(a) = \dfrac{l_t(a)}{p_t(a)} 1\{a = a_t\}, \forall a \in [K]$

1. Adversarial bandit

Algorithm - Exp3

- **Estimator** $\hat{l}_t \in \mathbb{R}_+^K$:
  - **Unbiased:** $\mathbb{E}[\hat{l}_t(a_t)] = p_t(a_t) \times \frac{l_t(a_t)}{p_t(a_t)} + (1 - p_t(a_t)) \times 0 = l_t(a_t)$
  - **Variance (2nd moment):**
    $\mathbb{E}[\hat{l}_t(a_t)^2] = p_t(a_t) \times \frac{l_t(a_t)}{p_t(a_t)^2}^2 + (1 - p_t(a_t)) \times 0^2 = \frac{l_t(a_t)^2}{p_t(a_t)} < \boxed{\frac{1}{p_t(a_t)}}$

Small $p_t(a_t)$ leads to high variance

1. Adversarial bandit

Algorithm - Exp3

- **Algorithm:**
  - Calculate $p_t(a) \propto \exp(-\eta \sum_{s<t} \hat{l}_s(a)), \forall a \in [K]$
  - Pick $a_t \sim p_t \in \Delta(K)$
  - Calculate $\hat{l}_t(a) = \frac{l_t(a)}{p_t(a)} 1\{a = a_t\}, \forall a \in [K]$

- **Exploitation:** Choose the action that has **smallest summation** of previous losses.

- **Exploration: Only** the loss of $a = a_t$ is non-zero, while the loss of others is **zero**. Therefore, the probability of seeing the **same action** next time is **decreased**. In other words, it encourages exploration.

  - **Remark:** l_t **cannot** be **negative!**

1. Adversarial bandit

Algorithm - Exp3

- **Regret bound:**
  - **Hedge**: $R_T = \mathcal{O}(\sqrt{T \ln K})$ with a condition $l_t \in [0,1]^K$
  - **Exp3:** $\hat{l}_t(a_t) = \dfrac{l_t(a_t)}{p_t(a_t)} < \dfrac{1}{p_t(a_t)}$, which can be large and the regret bound does not hold.
  - **Theorem 1:** With $\eta = \sqrt{\dfrac{\ln K}{TK}}$, Exp3 ensures $\mathbb{E}[R_T] = \mathcal{O}(\sqrt{T\boxed{K}\ln K})$

The price of knowing $\dfrac{1}{K}$ information

1. Adversarial bandit

Algorithm - Exp3

- **Regret bound:**
  - **Theorem 1:** With $\eta = \sqrt{\frac{\ln K}{TK}}$, Exp3 ensures $\mathbb{E}[R_T] = \mathcal{O}(\sqrt{TK \ln K})$
  - **Proof**:

$$
\begin{aligned}
\mathbb{E}[R_T] &\leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \mathbb{E}[\hat{l}_t(a)^2] \\
&= \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} \boxed{p_t(a) \frac{l_t(a)^2}{p_t(a)}} \longleftarrow \text{Variance cancellation} \\
&= \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} l_t(a)^2 \\
&\leq \frac{\ln K}{\eta} + \eta TK
\end{aligned}
$$

1. Adversarial bandit

Lower bound

- **Theorem**: For any learning algorithm, we can achieve $\Omega(\sqrt{TK})$
- **Notation**:
    - Loss of random action $l_t(a) \sim Ber(\frac{1}{2}), \forall a \neq a'$, Loss of "good" action $l_t(a') \sim Ber(\frac{1}{2} - \epsilon)$, with $\epsilon \in (0, \frac{1}{4}]$
    - $P_{a'}$: Probability conditioned on a'-th action is good (Similar for $\mathbb{E}_{a'}$)
    - P_{unf}: Probability with respect to uniformly random choice of loss for all actions (Similar for $\mathbb{E}_{unf}$)
    - a_t: Action at time t
    - l_t: loss at time t: Losses for all arms are generated independently and uniformly $\{0,1\}$, i.e $Ber(\frac{1}{2})$
    - n_a: # times action a is selected ($n_a = \sum_{t=1}^{T} 1(a_t = a)$)

1. Adversarial bandit

Lower bound

$$\mathbb{E}[l_t] = \mathbb{E}[l_t(a_t) \mid a_t = a']P(a_t = a') + \mathbb{E}[l_t(a_t) \mid a_t \neq a']P(a_t \neq a') = (\frac{1}{2} - \epsilon)P(a_t = a') + \frac{1}{2}P(a_t \neq a')$$

Proof:

$$E_{a'}[L_A - L_{min}] \geq \frac{1}{K}\sum_{a=1}^{K}\mathbb{E}_{a'}[L_A] - T(\frac{1}{2} - \epsilon) \qquad E_{a'}[L_{min}] \leq \sum_{t=1}^{T} E_{a'}[l_t(a_t = a')] = T(\frac{1}{2} - \epsilon)$$

$$= \frac{1}{K}\sum_{a'=1}^{K}\sum_{t=1}^{T}\mathbb{E}_{a'}[l_t] - T(\frac{1}{2} - \epsilon)$$

$$= \frac{1}{K}\sum_{a'=1}^{K}\sum_{t=1}^{T}[(\frac{1}{2} - \epsilon)P_{a'}(a_t = a') + \frac{1}{2}P_{a'}(a_t \neq a')] - T(\frac{1}{2} - \epsilon)$$

$$= \frac{1}{K}\sum_{a'=1}^{K}[\frac{T}{2} - \epsilon\mathbb{E}_{a'}[n_{a'}]] - T(\frac{1}{2} - \epsilon)$$

$$= T\epsilon - \frac{\epsilon}{K}\sum_{a'=1}^{K}\boxed{\mathbb{E}_{a'}[n_{a'}]}$$

Change from regret bound to action selection bound

1. Adversarial bandit

Lower bound

- **Proof (continue):** Let focus on the term $\sum_{a=1}^{K} \mathbb{E}_{a'}[n_{a'}]$.

$$\mathbb{E}_{a'}[n(a')] - E_{unf}[n(a')] = \sum_{l_{1:T}} n(a')(P_{a'}(l_{1:T}) - P_{unf}(l_{1:T}))$$

$$\leq T \sum_{l_{1:T}} P_{a'}(l_{1:T}) - P_{unf}(l_{1:T})$$

$$= T||P_{a'} - P_{unf}||_1$$

$$\leq T\sqrt{2\boxed{KL(P_{unf}||P_{a'})}}$$

Pinsker's inequality

Change from action selection bound to KL divergence

1. Adversarial bandit

Lower bound

$$\mathbb{E}[n_{a'}] = \mathbb{E}[\sum_{t=1}^{T} 1_{a_t=a'}] = \sum_{t=1}^{T} P(a_t = a')$$

- **Proof (continue):**

$$\mathbb{E}_{a'}[n_{a'}] - E_{unf}[n_{a'}] \leq \sqrt{2KL(P_{unf}||P_{a'})}$$

$$= T\sqrt{2\sum_{t=1}^{T} KL(P_{unf}(l_t|l_{1:t-1})||P_{a'}(l_t|l_{1:t-1}))}$$

Chain rule for KL divergence

$$= T\sqrt{2\sum_{t=1}^{T} P(a_t \neq a')KL(Ber(\frac{1}{2})||Ber(\frac{1}{2})) + P(a_t = a')KL(Ber(\frac{1}{2})||Ber(\frac{1}{2} - \epsilon))}$$

$$= T\sqrt{2\sum_{t=1}^{T} P(a_t = a')\frac{1}{2}\ln\frac{1}{1-4\epsilon^2}}$$

$$= T\sqrt{16\epsilon^2 \boxed{E_{unf}[n_{a'}]}}$$

$$\ln\frac{1}{1-x} \leq 4x, \forall x \leq \frac{1}{2}$$

Change KL divergence selection bound to action selection bound

1. Adversarial bandit

Lower bound

- **Proof (continue):**

$$\sum_{a'=1}^{K} \mathbb{E}_{a'}[n_{a'}] \leq \sum_{a'=1}^{K} E_{unf}[n_{a'}] + \sum_{a'=1}^{K} T\sqrt{16\epsilon^2 E_{unf}[n_{a'}]}$$

Adding $\sum_{a'=1}^{K}$ on both sides

$$\leq T + T\sqrt{16\epsilon^2 TK}$$

$\sum_{a'=1}^{K} E_{unf}[n_{a'}] = T,$ then $\sum_{a'=1}^{K} \sqrt{E_{unf}[n_{a'}]} \leq K\sqrt{\frac{T}{K}} = \sqrt{TK}$

$$= T + 4T\epsilon\sqrt{TK}$$

1. Adversarial bandit

Lower bound

- **Proof (continue)**: Then, we have

$$
\begin{aligned}
E_{a'}[L_A - L_{min}] &\geq T\epsilon - \frac{\epsilon}{K}\sum_{a'=1}^{K}\mathbb{E}_{a'}[n_{a'}] \\
&\geq T\epsilon - \epsilon(\frac{T}{K} + 4T\epsilon\sqrt{\frac{T}{K}}) \\
&= T\epsilon - \frac{T\epsilon}{K} - 4T\epsilon^2\sqrt{\frac{T}{K}} \\
&\geq \frac{T\epsilon}{2} - 4T\epsilon^2\sqrt{\frac{T}{K}} \\
&= \Omega(\sqrt{TK}) \qquad \text{Setting } \epsilon = \frac{1}{16}\sqrt{\frac{K}{T}}
\end{aligned}
$$

1. Adversarial bandit

Minimax optimal MAB - Intuition

- **Problem**: Gap $\sqrt{\ln K}$ between lower bound and upper bound.
- **Observation**: Hedge is FTRL with entropy regularizer. Exp3 is Hedge with loss estimator. Therefore, Exp3 is FTRL with entropy regularizer and loss estimator. Can we change the regularizer to close the gap?
- **Solution**: FTRL with **Tsallis** regularization function

1. Adversarial bandit

Minimax optimal MAB - Tsallis function

- **Tsallis function**: $\psi(p) = \dfrac{1 - \sum_{a=1}^{K} p(a)^{\beta}}{1-\beta}$

- **Remark**: Generalization of Shannon entropy function!

$$\lim_{\beta \to 1} \psi(p) = \lim_{\beta \to 1} \frac{1 - \sum_{a=1}^{K} p(a)^{\beta}}{1-\beta} = \lim_{\beta \to 1} \frac{-\sum_{a=1}^{K} p(a)^{\beta} \ln p(a)}{-1} = \sum_{a=1}^{K} p(a) \ln p(a)$$

1. Adversarial bandit

Minimax optimal MAB - Algorithm

- **FTRL**: $p_{t+1} = \arg\min_{p \in \Delta(K)} \langle p, \sum_{s<t+1} \hat{l}_s \rangle + \frac{1}{\eta}\psi(p)$
  - Use OMD framework to solve for p_{t+1} and obtain regret bound
  - **Remark**: FTRL framework is possible, but more difficult!

1. Adversarial bandit

Minimax optimal MAB - OMD

$$\psi(p) = \frac{1 - \sum_{a=1}^{K} p(a)^\beta}{1 - \beta}$$

- **OMD framework**

(1) $\nabla \psi(p'_{t+1}) = \nabla \psi(p_t) - \eta \hat{l}_t \longrightarrow \frac{1}{p'_{t+1}(a)^{1-\beta}} = \frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta} \eta \hat{l}_t(a)$

(2) $$\begin{aligned} p_{t+1} &= \arg\min_{p \in \Delta(K)} D_\psi(p, p'_{t+1}) \longrightarrow \frac{1}{p_{t+1}(a)^{1-\beta}} = \frac{1}{p'_{t+1}(a)^{1-\beta}} + \lambda \\ &= \arg\min_{p \in \Delta(K)} \psi(p) - \psi(p'_{t+1}) - \langle \nabla \psi(p'_{t+1}), p - p'_{t+1} \rangle \\ &= \arg\min_{p \in \Delta(K)} \frac{1}{1-\beta} \sum_{a=1}^{K} \left(p'_{t+1}(a)^\beta - p(a)^\beta + \frac{\beta}{p'_{t+1}(a)^{1-\beta}}(p(a) - p'_{t+1}(a))\right) \end{aligned}$$

Lagrange multiplier

1. Adversarial bandit

Minimax optimal MAB - Update rule

- **FTRL**: $p_{t+1} = \arg\min_{p \in \Delta(K)} \langle p, \sum_{s<t+1} \hat{l}_s \rangle + \frac{1}{\eta}\psi(p)$
  - $\frac{1}{p'_{t+1}(a)^{1-\beta}} = \frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta}\eta \hat{l}_t(a)$

    $\frac{1}{p_{t+1}(a)^{1-\beta}} = \frac{1}{p'_{t+1}(a)^{1-\beta}} + \lambda$

    $= \frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta}\eta \hat{l}_t(a) + +\lambda$

    $= \frac{1}{p'_t(a)^{1-\beta}} + \frac{1-\beta}{\beta}\eta \hat{l}_t(a) + +2\lambda$ **(Recursion)**
  - $= \dots$

    $= \frac{1-\beta}{\beta}\eta(\sum_{s<t+1} \hat{l}_t(a)) + \lambda'$

Update rule: $\frac{1}{p_{t+1}(a)^{1-\beta}} = \frac{1-\beta}{\beta}(\lambda' + \eta \sum_{s<t+1} \hat{l}_t(a))$

1. Adversarial bandit

Minimax optimal MAB - Regret

- For $q \in \Delta(K)$

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \langle p_t, \hat{l}_t \rangle - \sum_{t=1}^{T} \langle q, \hat{l}_t \rangle \\
&= \sum_{t=1}^{T} \langle p_t - q, \hat{l}_t \rangle \\
&= \frac{1}{\eta} \sum_{t=1}^{T} D_\psi(q, p_t) - D_\psi(q, p'_{t+1}) + D_\psi(p_t, p'_{t+1}) && \text{Three-point inequality} \\
&\le \frac{1}{\eta} \sum_{t=1}^{T} D_\psi(q, p_t) - D_\psi(q, p_{t+1}) + D_\psi(p_t, p'_{t+1}) && \text{Definition of } p_{t+1} \\
&= \frac{D_\psi(q, p_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} D_\psi(p_t, p'_{t+1}) \\
&= \frac{K^{1-\beta} - 1}{1 - \beta} + \frac{1}{\eta} \sum_{t=1}^{T} D_\psi(p_t, p'_{t+1}) && \text{Range of } \psi(p)
\end{aligned}
$$

1. Adversarial bandit

Minimax optimal MAB - Regret

- For $q \in \Delta(K)$

$$R_T = \frac{K^{1-\beta}-1}{1-\beta} + \frac{1}{\eta}\sum_{t=1}^{T}\sum_{a=1}^{K} p'_{t+1}(a)^{\beta} - p_t(a)^{\beta} + \eta p_t(a)\hat{l}_t(a)$$

OMD framework

$$= \frac{K^{1-\beta}-1}{1-\beta} + \frac{1}{\eta}\sum_{t=1}^{T}\sum_{a=1}^{K}\left(p_t(a)^{\beta}\left(1+\frac{1-\beta}{\beta}\eta p_t(a)^{1-\beta}\hat{l}_t(a)\right)^{\frac{\beta}{\beta-1}} - p_t(a)^{\beta} + \eta p_t(a)\hat{l}_t(a)\right)$$

$$= \frac{K^{1-\beta}-1}{1-\beta} + \frac{1}{\eta}\sum_{t=1}^{T}\sum_{a=1}^{K}\left(p_t(a)^{\beta}\left(1-\eta p_t(a)^{1-\beta}\hat{l}_t(a)+\frac{\eta^2}{\beta}p_t(a)^{2-2\beta}\hat{l}_t(a)^2\right) - p_t(a)^{\beta} + \eta p_t(a)\hat{l}_t(a)\right)$$

$(1+x)^{\alpha} \leq 1+\alpha x+\alpha(\alpha-1)x^2 \forall x \geq 0$ and $\alpha < 0$

$$= \frac{K^{1-\beta}-1}{1-\beta} + \frac{\eta}{\beta}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^{2-\beta}\hat{l}_t(a)^2$$

1. Adversarial bandit

Minimax optimal MAB - Regret

- **Regret bound**:

$$E[R_T] \le \frac{K^{1-\beta}-1}{1-\beta} + \frac{\eta}{\beta}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^{1-\beta}$$

$$\le \frac{K^{1-\beta}-1}{1-\beta} + \frac{\eta}{\beta}\sum_{t=1}^{T}(\sum_{a=1}^{K}(p_t(a)^{1-\beta})^{\frac{1}{1-\beta}})^{1-\beta}(\sum_{a=1}^{K}1^{\frac{1}{\beta}})^{\beta}$$

Holder's inequality

$$= \frac{K^{1-\beta}-1}{1-\beta} + \frac{\eta}{\beta}TK^{\beta}$$

- $\beta = \frac{1}{2}, \eta = \frac{1}{\sqrt{T}}$, then $E[R_T] \le 4\sqrt{TK}$

1. Adversarial bandit

Summary

- **MAB**: Partial information feedback
- **EXP3**: Hedge with unbiased, but large variance, loss estimator
    - **Regret of EXP3**: $\Omega(\sqrt{TK}), \mathcal{O}(\sqrt{TK \ln K})$
- **Minimax optimal MAB**: FTRL with Tsallis regularizer, $\mathcal{O}(\sqrt{TK})$

2. Stochastic Bandit

Setting - Notation

- **Action at time t**: $a_t \in [K]$
- **Loss at time t:** $l_t \in [0,1]^K$
    - **Remark**: Only know $l_t(a_t)$
- **Setting**: Stochastic
    - Each arm has unknown loss distribution D_a and $\mu(a)$
    - I.i.d: Does not depend on t
    - **Example**: D_a are Bernoulli distribution (0/1 loss)

2. Stochastic Bandit

Setting - Regret

Regret: $\mathbb{E}[R_T] = \mathbb{E}[\sum_{t=1}^{T} l_t(a_t)] - \min_{a \in [K]} \sum_{t=1}^{T} l_t(a)$

- **Pseudo-Regret**: $\bar{R}_T = \max_{a \in K} \mathbb{E}[\sum_{t=1}^{T} l_t(a_t) - \sum_{t=1}^{T} l_t(a)]$
    - **Stochasticity**: Algorithm choice and environment
    - **Remark:** Pseudo-regret is always smaller than expected regret
- **Another version:** $\bar{R}_T = \mathbb{E}[\sum_{t=1}^{T} (\mu(a_t) - \mu(a^*)]$ with $a^* = \arg\min_a \mu(a)$
    - **Remark:** Possible to **ignore the deviation** between $l_t(a)$ and $\mu(a)$

2. Stochastic Bandit

Setting - Non-adaptive vs adaptive algorithm

- Trade-off between exploration and exploitation!
- **Non-adaptive versus adaptive**: Previous experience to guide exploration
- **Non-adaptive algorithm:** Explore-then-Exploit and ϵ-greedy
- **Adaptive algorithm:** Successive elimination and UCB

2. Stochastic Bandit

Setting - Lemma

Radius r: key component for every proof!

- **Lemma**: For stochastic bandit setting, no matter the learning strategy is, we have that $\Pr(|\hat{\mu}_t(a) - \mu(a)| \leq 2\boxed{\sqrt{\frac{\log T}{n_t(a)}}}) \geq \boxed{1 - \frac{2K}{T}}$ ← High probability event

- **Intuition:** Among T rounds, if $n_t(a) \approx T$, then $\hat{\mu}_t(a) \approx \mu(a)$ with high probability

- **Proof**: Hoeffding's inequality

2. Stochastic Bandit

Algorithm - Explore-then-Exploit

$$\Pr(|\hat{\mu}(a) - \mu(a)| \leq 2\sqrt{\frac{\log T}{n_t(a)}}) \geq 1 - \frac{2K}{T}$$

- **Intuition:** Try each arms multiple times to estimate $\hat{\mu}(a)$ for every a. Then, repeatedly select the lowest estimated arm.
- **Algorithm**: Among T rounds, **select at least N** rounds for each arm. Each arm will have $\hat{\mu}(a) = \frac{1}{N}\sum_{t=1}^{N} l_t(a)$. Then, choose the arm with l**owest estimation** for the rest $T - NK$ rounds.

2. Stochastic Bandit

$$\Pr(|\hat{\mu}(a) - \mu(a)| \leq 2\sqrt{\frac{\log T}{n_t(a)}}) \geq 1 - \frac{2K}{T}$$

Algorithm - Explore-then-Exploit - Regret analysis

- Assume we have $K = 2$, a and a^*, $n_t(a) = n_t(a^*) = N$
- $\mu(a) \in [\hat{\mu}_t(a) - r_t(a), \hat{\mu}_t(a) + r_t(a)]$ with $r_t(a) = 2\sqrt{\frac{\log T}{n_t(a)}}$
- If we choose a^* instead of a (**exploitation**), then

$$\mu(a^*) - r_t(a^*) < \hat{\mu}_t(a^*) < \hat{\mu}_t(a) < \mu(a) + r_t(a)$$

- In other words, $\mu(a^*) - \mu(a) \leq 4\sqrt{\frac{\log T}{N}}$
  - **Remark**: **Not bad** if we select **non-optimal** arm with large N

2. Stochastic Bandit

$$\Pr(|\hat{\mu}(a) - \mu(a)| \leq 2\sqrt{\frac{\log T}{n_t(a)}}) \geq 1 - \frac{2K}{T}$$

Algorithm - Explore-then-Exploit - Regret analysis

- **Exploitation:** $\Delta_a = \mu(a^*) - \mu(a) \leq 4\sqrt{\frac{\log T}{N}}$

Exploration

- **Regret (2 arms):** $R(T) \leq \boxed{2N} + 4(T-2N)\sqrt{\frac{\log T}{N}}$.
    - With $N = T^{2/3}(\log T)^{2/3}$, then $R(T) \leq \mathcal{O}(T^{2/3}(\log T)^{1/3})$

$$\begin{aligned}\bar{R}_T = \mathbb{E}[R(T)] &= \mathbb{E}[R(T) \mid \text{Good event}]\Pr(\text{Good event}) + \mathbb{E}[R(T) \mid \text{Bad event}]\Pr(\text{Bad event}) \\ &\leq \mathbb{E}[R(T) \mid \text{Good event}] + T \times O(T^{-1}) \\ &\leq \mathcal{O}(T^{2/3}(\log T)^{1/3})\end{aligned}$$

Worse than $\sqrt{T}$

- **Generalized regret:** $\bar{R}_T \leq \mathcal{O}(\boxed{T^{2/3}}(K \log T)^{1/3})$ - Every-instance regret
    - **Remark**: Worse than the bound obtained by EXP3 in MAB $\mathcal{O}(\sqrt{TK \log K})$

2. Stochastic Bandit

Algorithm - Explore-then-Exploit - Regret analysis

- **Exploitation**: $\Delta_a = \mu(a^*) - \mu(a) \leq 4\sqrt{\frac{\log T}{N}} \rightarrow N \leq 16\frac{\log T}{\Delta_a^2}$ (Change the order)
- **Generalized regret:** $R(T) \leq \sum_a 16\frac{\log T}{\Delta_a^2} + \sum_a T\Delta_a$.
    - **Remark:** If Δ_a is **small**, then Δ_a^2 is even **smaller** and $R(T)$ is **loose**.

2. Stochastic Bandit

Algorithm - ϵ-Greedy

- **Motivation:** If arm have large Δ, then a lot of **redundant** exploration.
- **Solution**: **Spread the exploration** more uniformly over time.
- **Algorithm:**

for *each round* $t = 1, 2, \ldots$ **do**
  Toss a coin with success probability $\boxed{\epsilon_t;}$ ← Change over time!
  if *success* **then**
    explore: choose an arm uniformly at random
  else
    exploit: choose the arm with the lowest average loss so far
end

2. Stochastic Bandit

Algorithm - ϵ-Greedy - Regret Analysis

for *each round* $t = 1, 2, \ldots$ **do**
  Toss a coin with success probability ϵ_t;
  if *success* **then**
    explore: choose an arm uniformly at random
  else
    exploit: choose the arm with the lowest average loss so far
end

- **Intuition:** ϵ_t - **Decreasing** sequence. **Less** exploration per arm!
- **Regret:** With $\epsilon_t = t^{-1/3}(K \log t)^{1/3}$, then $\mathbb{E}[R(T)] \leq \mathcal{O}(t^{2/3}(K \log t)^{1/3})$
    - **Remark:** This regret bound is **stronger** since it holds for **all** rounds $t \in [T]$

2. Stochastic Bandit

Algorithm - Successive elimination

- **Motivation**: Previous algorithms are non-adaptive since they do not use **previous experience** to prevent **useless exploration**!
- **Solution:** Use previous experience to **eliminate the useless arms** as soon as possible!
- **Algorithm**:

```
All arms are initially designated as active
loop {new phase}
   play each active arm once
   deactivate all arms a' such that, letting t be the current round,
      UCB_t(a) < LCB_t(a') for some other arm a' {deactivation rule}
end loop
```

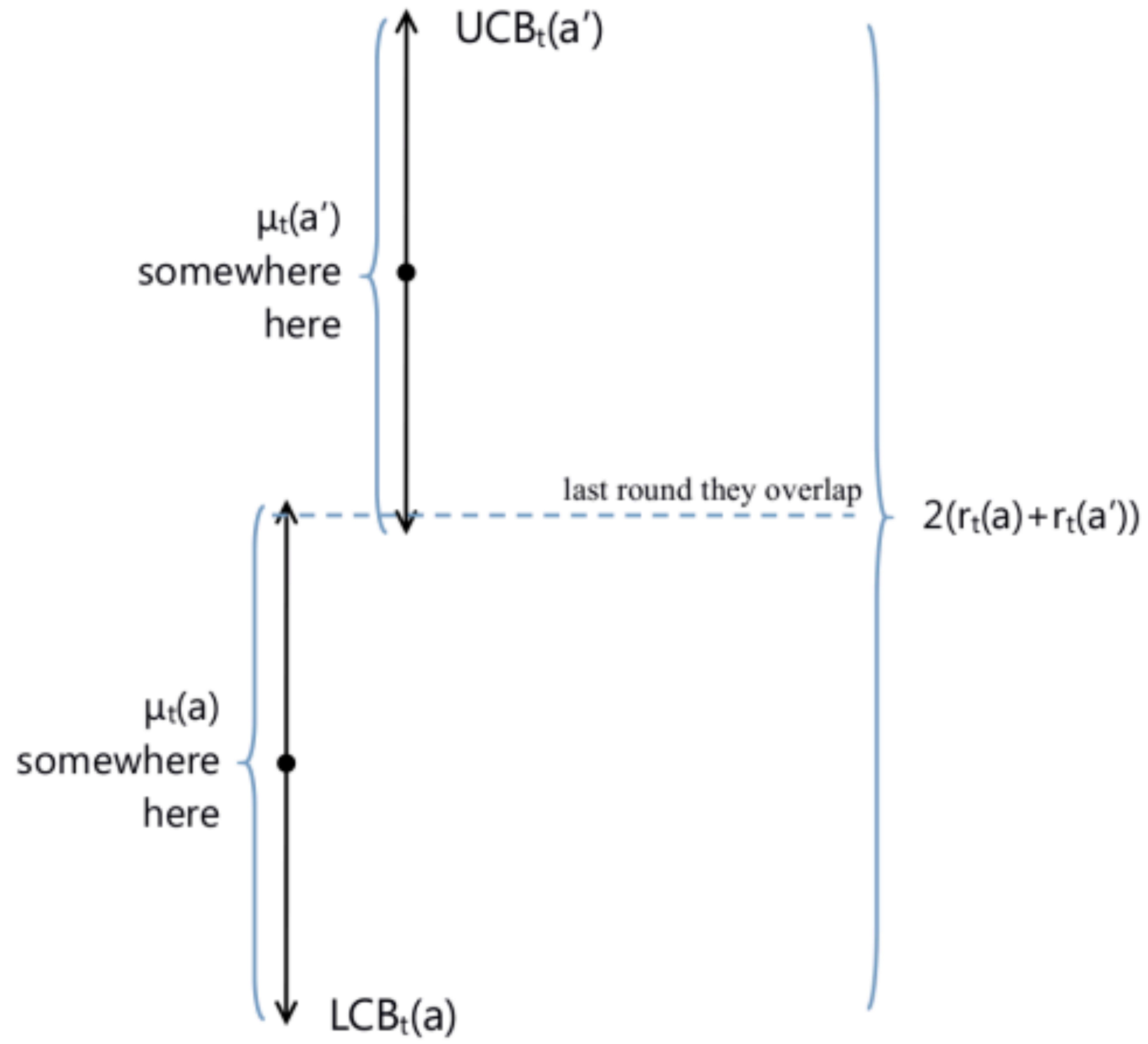
2. Stochastic Bandit

Algorithm - Successive elimination

$$LCB_t(a) = \hat{\mu}_t(a) - r_t(a)$$
$$UCB_t(a) = \hat{\mu}_t(a) + r_t(a)$$

All arms are initially designated as *active*
loop {new phase}
  play each active arm once
  deactivate all arms a' such that, letting t be the current round,
    $\mathtt{UCB}_t(a) < \mathtt{LCB}_t(a')$ for some other arm a' {deactivation rule}
end loop

- **Intuition**: Since $\mu(a) \in \hat{\mu}_t(a) \pm r_t(a)$, then $UCB(a) < LCB(a')$ means $\mu(a) < \mu(a')$ with high probability.



2. Stochastic Bandit

$$R_T = \sum_{t=1}^{T} (\mu(a_t) - \mu(a^*))$$

Algorithm - Successive elimination - Regret analysis

- $\boxed{\Delta_a = \mu(a^*) - \mu(a) \leq 2(\sqrt{\frac{\log T}{n_t(a)}} + \sqrt{\frac{\log T}{n_t(a^*)}}) \leq 4\sqrt{\frac{\log T}{n_t(a)}}}$ ← Key equation!

  Hint: Previous pic
- $R_T(a) = n_T(a)\Delta_a \leq 4\sqrt{n_T(a)\log T}$
- $R_T = \sum_a R_T(a) \leq \sum_a 4\sqrt{n_T(a)\log T} \leq \mathcal{O}(\sqrt{KT\log T})$ where the last inequality is obtained by Cauchy-Schwarz's inequality $\frac{1}{n}\sum_{i=1}^{n} f(x_i) \leq f(\frac{1}{n}\sum_{i=1}^{n} x_i)$
- $\bar{R}_T = \mathbb{E}[R_T] = \mathcal{O}(\sqrt{TK\log T})$ - This is every-instance bound!

2. Stochastic Bandit

$$R_T = \sum_{t=1}^{T} (\mu(a_t) - \mu(a^*))$$

Algorithm - Successive elimination - Regret analysis

- $\Delta_a \leq 4\sqrt{\frac{\log T}{n_t(a)}} \rightarrow n_t(a) \leq 16\frac{logT}{\Delta_a^2}$ (Change the order)
- $R_T(a) = n_T(a)\Delta_a \leq 16\frac{\log T}{\Delta_a}$
- $R_T = \sum_a R_T(a) = \sum_a 16\frac{\log T}{\Delta_a} \leq \mathcal{O}(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a})$ **-** instance-dependence bound!
- **Intuition**: If all arms are **similar**, which means it is **hard to distinguish** the best arm, then Δ_a will be **small** and regret bound will be **large**!
- **Remark:** but for the worst-case instance, the maximum regret is $\mathcal{O}(\sqrt{TK\log T})$

2. Stochastic Bandit

$$LCB_t(a) = \hat{\mu}_t(a) - r_t(a)$$

Algorithm - Upper Confidence Bound (UCB)

- **Motivation**: Assume each arm is **as good as** it possibly be given the **observations so far**, and choose the best arm based on the **optimistic** estimates!

- **Algorithm:**

> Try each arm once
> **for** each round $t = 1, \ldots, T$ **do**
> pick arm some a which minimizes LCB$_t(a)$.
> **end for**

- **Intuition: Small** LCB means either **small** $\hat{\mu}$ (exploit) or **large** r_t (explore)

2. Stochastic Bandit

Algorithm - Upper Confidence Bound (UCB) - Regret analysis

Try each arm once
for each round $t = 1, \ldots, T$ **do**
pick arm some a which minimizes LCB$_t(a)$.
end for

- Assume there are **2 arms** and the algorithm **choose non-optimal** arm a_t.
- $\hat{\mu}_t(a_t) - r_t(a_t) = LCB(a_t) < LCB(a^*) < \mu(a^*)$ and $\mu(a_t) < \hat{\mu}_t(a_t) + r_t(a_t)$.
- Then, $\boxed{\Delta_{a_t} = \mu(a_t) - \mu(a^*) \leq 2r_t(a_t) = 4\sqrt{\frac{\log T}{n_t(a_t)}}}$ Key equation appears **again**!
- $\bar{R}_T = \mathcal{O}(\sqrt{TK \log T})$

2. Stochastic Bandit

Algorithm - Lower bound

- **Non-adaptive**:
  - **Every-instance bound**: $\boxed{\Omega(T^{2/3}K^{1/3})}$ ← Less than $(\log T)^{1/3}$
  - **Instance-dependent bound**: $\boxed{\Omega(C^{-2}T^{\lambda}\sum_a \Delta(a))}$ if $\bar{R}_T \leq C \cdot T^{\gamma}$, $\gamma \in [2/3,1)$, $C > 0$ and $\lambda = 2(1-\gamma)$ Less than $\sum_a \frac{1}{\Delta(a)^2}$
  - **Remark:** With $\sum_a \Delta(a)$ sufficiently small, $C = K^{-1/6}$, and $\lambda = 2/3$, we have two above bound are **equivalent**!

2. Stochastic Bandit
Algorithm - Lower bound

- **Non-adaptive (Proof):**
    -

2. Stochastic Bandit

Algorithm - Lower bound

- **Adaptive:**
  - **Every-instance bound:** $\Omega(\sqrt{KT})$ - Previous result!
  - **Instance-dependent bound:** Define a problem instance $I = \{\mu(a)\,|\,a \in K\}$ and any algorithms that satisfies $\bar{R}_t \leq C_{I,\alpha} t^{\alpha}, \quad \forall \alpha > 0$. Then, there exists t_0 such that $\forall t \geq t_0$, we have $\bar{R}_t \geq C_I \ln t$ with

$\mu^*(1-\mu^*) < 1$

  - $C_I = \begin{cases} \sum_{a:\Delta(a)>0} \dfrac{\boxed{\mu^*(1-\mu^*)}}{\Delta(a)} & \text{(Easier case)} \\ \sum_{a:\Delta(a)>0} \dfrac{\Delta(a)}{KL(\hat{\mu}_a \,||\, \mu_{a^*})} - 2\epsilon \quad \forall \epsilon > 0 & \text{(Hard case)} \end{cases}$

2. Stochastic Bandit
Algorithm - Lower bound

- **Adaptive: (Proof)**
  - **Every-instance bound:** $\Omega(\sqrt{KT})$ - Previous result!
  - **Instance-dependent bound:** Define a problem instance $I = \{\mu(a) \,|\, a \in K\}$ and any algorithms that satisfies $\bar{R}_t \leq C_{I,\alpha} t^{\alpha}, \quad \forall \alpha > 0$. Then, there exists t_0 such that $\forall t \geq t_0$, we have $\bar{R}_t \geq C_I \ln t$ with

$$\mu^*(1-\mu^*) < 1$$

  - $$C_I = \begin{cases} \sum_{a:\Delta(a)>0} \frac{\boxed{\mu^*(1-\mu^*)}}{\Delta(a)} & \text{(Easier case)} \\ \sum_{a:\Delta(a)>0} \frac{\Delta(a)}{KL(\hat{\mu}_a || \mu_{a^*})} - 2\epsilon \quad \forall \epsilon > 0 & \text{(Hard case)} \end{cases}$$

2. Stochastic Bandit

Algorithm - Lower bound

- **Adaptive: (Proof)**
  -

2. Stochastic Bandit

Summary

- **Stochastic environment**: unknown D_a and μ_a, Pseudo-regret
- **Non-adaptive:** Explore-then-Exploit, $\epsilon-$greedy
- **Adaptive:** Successive elimination, UCB

3. Bayesian Bandit

Setting - Motivation

- Previous section shows 2 versions of regret: **instance-dependent** and **all-instance**
- **Instance-dependent regret:** regret is being affected largely by pre-defined instance I.
- **Question:** What if we pre-define the distribution of I, sample and then update the distribution accordingly?
- **Answer:** Bayesian bandit!

3. Bayesian Bandit

Setting - Notation

- **Prior instance**: $I = \{\mu(a) \,|\, a \in K\}$ as well as $\mathscr{D}_a$
- **t-history**: $H_t = ((a_1, l_1), \ldots, (a_t, l_t)) \in (\mathscr{A} \times \mathbb{R})^t$
- **Feasible t-history**: $H = ((a'_1, l'_1), \ldots, (a'_t, l'_t)) \in (\mathscr{A} \times \mathbb{R})^t$
- **H-consistent algorithm**: $\Pr(H_t = H) > 0$
- **H-induced algorithm**: $a_s = a'_s, \forall s \in [t]$ (deterministically)

3. Bayesian Bandit

Setting - Notation

$$\bar{R}_T = \mathbb{E}[\sum_{t=1}^{T} (\mu(a_t) - \mu(a^*)]$$

- **Posterior distribution of $\mathbb{P}$ given H**:
  $\mathbb{P}_H(\mathscr{M}) = \Pr(\mu \in \mathscr{M} \mid H_t = H), \forall \mathscr{M} \subset [0,1]^K$

- **Bayesian Regret**: $BR(T) = \mathbb{E}_{I \sim \mathbb{P}}[\sum_{t=1}^{T} \mu(a_t) - T\mu^*]$

3. Bayesian Bandit

Setting - Notation

- **Lemma**: Distribution $\mathbb{P}_H$ is the same for all H-consistent bandit algorithm. In other words, $\mathbb{P}_H$ does not depend on which H-consistent bandit algorithm has collected the history.
- Suppose $\mathscr{M} = \{\tilde{\mu}\}$, then $P_H(\tilde{\mu}) = \dfrac{\mathbb{P}_{H'}(\tilde{u}) \cdot \mathscr{D}_{\tilde{\mu}(a)}(l)}{\sum_{\tilde{\mu}} \mathbb{P}_{H'}(\tilde{u}) \cdot \mathscr{D}_{\tilde{\mu}(a)}(l)}$, where
    - $\mathbb{P}_H(\tilde{\mu}) = P(\mu = \tilde{\mu} \,|\, H_t = H)$
    - $\mathscr{D}_{\tilde{\mu}(a)}(l)$ is probability of receive loss l with loss distribution with mean loss $\tilde{\mu}(a)$.
    - H': is the previous history of H_{t-1}

3. Bayesian Bandit

Setting - Notation

- **Proof:** $P_H(\tilde{\mu}) = \dfrac{\Pr(\mu = \tilde{\mu} \text{ and } H_t = H)}{\Pr(H_t = H)}$

- $\Pr(\mu = \tilde{\mu} \text{ and } H_t = H) = \Pr(\mu = \tilde{\mu})\mathscr{D}_{\tilde{\mu}(a)}(l)\pi(a)\Pr(H_{t-1} = H')$ where
    - $\mathscr{D}_{\tilde{\mu}(a)}(l) = \Pr(l_t = l \,|\, a_t = a \text{ and } \mu = \tilde{\mu} \text{ and } H_{t-1} = H')$
    - $\pi(a) = \Pr(a_t = a \,|\, H_{t-1} = H')$

No $\pi(a)$

- $P_H(\tilde{\mu}) = \dfrac{\Pr(\mu = \tilde{\mu} \text{ and } H_t = H)}{\Pr(H_t = H)} = \dfrac{\Pr(\mu = \tilde{\mu} \text{ and } H_t = H)}{\sum_{\tilde{\mu}} \Pr(\mu = \tilde{\mu} \text{ and } H_t = H)} = \boxed{\dfrac{\mathbb{P}_{H'}(\tilde{u}) \cdot \mathscr{D}_{\tilde{\mu}(a)}(l)}{\sum_{\tilde{\mu}} \mathbb{P}_{H'}(\tilde{u}) \cdot \mathscr{D}_{\tilde{\mu}(a)}(l)}}$

3. Bayesian Bandit

Setting - Notation

- **Posterior as a new prior**:
  $\mathbb{P}_{H \oplus H'}(\mathcal{M}) = P_{\mu \sim \mathcal{P}_H}(\mu \in \mathcal{M} \mid H_{t'} = H') = P_{\mu \sim \mathcal{P}_H}(\mu \in \mathcal{M} \mid H_{t'} = H', H_t = H)$

- **Intuition:** Previous information H has encompassed, which means we can forget about past interaction!

- **Proof**: Very similar to the previous proof!

3. Bayesian Bandit

Algorithm - Thompson Sampling

Thompson Sampling admits an alternative characterization:

for *each round* $t = 1, 2, \ldots$ **do**
  Observe $H_{t-1} = H$, for some feasible $(t-1)$-history H;
  Sample mean reward vector μ_t from the posterior distribution $\mathbb{P}_H$;
  Choose the best arm $\tilde{a}_t$ according to μ_t.
end

Algorithm 3.2: Thompson Sampling: alternative characterization.

- $\mathbb{P}_H(\tilde{\mu}) = \dfrac{P(H_t = H \mid \mu = \tilde{\mu})\mathbb{P}(\tilde{\mu})}{\sum_{\tilde{\mu}} P(H_t = H \mid \mu = \tilde{\mu})\mathbb{P}(\tilde{\mu})}$

- **Problem:** Mathematically well-defined, but computationally inefficiency!
  - Computation of $P(H_t = H \mid \mu = \tilde{\mu})$ is t and the total computation of $\mathbb{P}_H$ is $t \cdot |\mathscr{F}|$

3. Bayesian Bandit

Algorithm - Thompson Sampling

Thompson Sampling admits an alternative characterization:

for *each round* $t = 1, 2, \ldots$ **do**

Observe $H_{t-1} = H$, for some feasible $(t-1)$-history H;

Sample mean reward vector μ_t from the posterior distribution $\mathbb{P}_H$;

Choose the best arm $\tilde{a}_t$ according to μ_t.

end

Algorithm 3.2: Thompson Sampling: alternative characterization.

- **Improvement**: sequential Bayesian update $P_{H'}(\tilde{\mu}) = \dfrac{\mathbb{P}_H(\tilde{u}) \cdot \mathcal{D}_{\tilde{\mu}(a)}(l)}{\sum_{\tilde{\mu}} \mathbb{P}_H(\tilde{u}) \cdot \mathcal{D}_{\tilde{\mu}(a)}(l)}$
- **Even faster**: Independent prior (independent arms), conjugate prior (Beta-Bernoulli, Guassian-Gaussian)

3. Bayesian Bandit

Algorithm - Regret

- **Lemma**: If $\mathbb{E}[[U(a,H_t) - \mu(a)]^-] \leq \frac{\gamma}{TK}$ and $\mathbb{E}[[\mu(a) - L(a,H_t)]^-] \leq \frac{\gamma}{TK}$, then $BR(T) \leq 2\gamma + 2\sum_{t=1}^{T}\mathbb{E}[l(a_t, H_t)]$ with $l(a_t, H_t) = U(a_t, H_t) - L(a_t, H_t)$
  - **Remark:** Do not depend the structure of prior, or specify how to calculate how to calculate U and L
- **Theorem**: With radius $l(a_t, H_t) = \sqrt{\frac{\log T}{n_t(a)}}$ define as above, then $\boxed{BR(T) \leq O(\sqrt{KT \log T})}$

Similar to previous result

4. Contextual Bandit

Revisted EXP3

- **Theorem**: Let $F_t = \sum_a p_t(a) \cdot c_t(a) = \mathbb{E}[c_t(a_t) \mid w_t]$ where $p_t(a) = \frac{w_t(a)}{\sum_a w_t(a)}$, $G_t = \sum_a p_t(a) \cdot c_t(a)^2 = \mathbb{E}[c_t(a_t)^2 \mid w_t]$ where $\sum_t \mathbb{E}[G_t] \leq uT$ for some known u. Then, $\alpha = \ln(\frac{1}{1-\epsilon}), \beta = \alpha^2, \epsilon = \sqrt{\frac{\ln K}{3uT}}$,

$$\mathbb{E}[cost(ALG) - cost^*] < 2\sqrt{3}\sqrt{uT \ln K}$$

4. Contextual Bandit

Revisted Exp3

- **Remark**: Exp3 does not have experts, only arms (no context). Exp3 is a generalization of Hedge with lost estimator
- **Motivation** Exp3 does not work well with correlated/dependent arms, large # arms leads to large regret!
- **Answer**: Exp4 - generalization of Exp3
- **Exp3** = Exponential-weight algorithm for Exploration and Exploitation
- **Exp4** = Exponential-weight algorithm for Exploration and Exploitation using **Expert Advices**

4. Contextual Bandit

Notation

- **Assumption**: $\mathbb{E}[\hat{c}_t(e) \mid \mathbf{p}_t] = c_t(e)$
- **Distribution over experts**: $p_t(e) := \Pr(e_t = e)$
- **Distribution over arms**: $q_t(a) = \Pr(a_t = a \mid p_t)$
- **Fake cost**: $\hat{c}_t(a) = \dfrac{c_t(a_t)}{q_t(a_t)}$ if $a_t = a$ and $\hat{c}_t(e) = \hat{c}_t(a_{t,e})$
- $\mathbb{E}[\hat{c}_t(a) \mid p_t] = c_t(a)$

4. Contextual Bandit

Notation

- **Regret (refined)**: $R(T) = cost(ALG) - \min_{e \in \mathscr{E}} cost(e)$ where $cost(e) = \sum_t c_t(a_{t,e}) = \sum_t c_t(e)$
  - **Another version**:
    $$R(T) = cost(ALG) - \sum_{t=1}^{T} \sum_{a=1}^{N} e_{t,i*}(a) l_t(a) = cost(ALG) - \min_i \sum_{t=1}^{T} \sum_{a=1}^{N} e_{t,i}(a) l_t(a)$$
  - **Remark**: Before, we compare the cost of the algorithm versus the best arm, which is unchanged throughout T rounds. However, the "best" experts can change their decision throughout T rounds, which is more flexible notion!

4. Contextual Bandit

Algorithm

Given: set $\mathcal{E}$ of experts, parameter $\epsilon \in (0, \frac{1}{2})$ for `Hedge`, exploration parameter $\gamma \in [0, \frac{1}{2})$.
In each round t,

1. Call `Hedge`, receive the probability distribution p_t over $\mathcal{E}$.
2. Draw an expert e_t independently from p_t.
3. *Selection rule*: with probability $1 - \gamma$ follow expert e_t; else pick an arm a_t uniformly at random.
4. Observe the cost $c_t(a_t)$ of the chosen arm.
5. Define fake costs for all experts e:

$$\widehat{c}_t(e) = \begin{cases} \frac{c_t(a_t)}{\Pr[a_t = a_{t,e} | \vec{p}_t]} & a_t = a_{t,e}, \\ 0 & \text{otherwise.} \end{cases}$$

6. Return the "fake costs" $\widehat{c}(\cdot)$ to `Hedge`.

4. Contextual Bandit

Algorithm

Algorithm 1 EXP4 for contextual bandits

Initialize $w_1 = (1, 1, \ldots, 1)$;
for $t = 1 \to T$ **do**
EG gives us probability over experts $p_t \in \Delta(M)$: $p_t = \frac{w_t}{\|w_t\|}$;
Compute probability q_t over actions by integrating out expert i: $\forall a, q_t(a) = \sum_{i=1}^{M} p_{t,i} e_{t,i}(a)$;
Draw action $a_t \sim q_t$, and incur loss $l_t(a_t)$;
Build the unbiased estimate for full feedback l_t:

$$\forall a, \hat{l}_t(a) = \begin{cases} l_t(a_t)/q_t(a_t), & \text{if } a = a_t; \\ 0, & \text{otherwise;} \end{cases}$$

Compute the expected loss g_t:

$$\forall i, g_{t,i} = \sum_{a=1}^{A} e_{t,i}(a) l_t(a) = \frac{e_{t,i}(a_t) l_t(a_t)}{q_t(a_t)};$$

EG update w_t from g_t: $\forall i, w_{t+1,i} = w_t \exp\left(-\eta \sum_{s=1}^{t} g_{t,i}\right)$;
end for

4. Contextual Bandit

Regret Analysis

- **Regret:** $\mathbb{E}[R(T)] \leq O(\sqrt{KT \log N})$
- **Proof:**

Proof. For each arm a, let $\mathcal{E}_a = \{e \in \mathcal{E} : a_{t,e} = a\}$ be the set of all experts that recommended this arm. Let

$$p_t(a) := \sum_{e \in \mathcal{E}_a} p_t(e)$$

be the probability that the expert chosen by `Hedge` recommends arm a. Then

$$q_t(a) = p_t(a)(1-\gamma) + \frac{\gamma}{K} \geq (1-\gamma)\, p_t(a).$$

For each expert e, letting $a = a_{t,e}$ be the recommended arm, we have:

$$\widehat{c}_t(e) = \widehat{c}_t(a) \leq \frac{c_t(a)}{q_t(a)} \leq \frac{1}{q_t(a)} \leq \frac{1}{(1-\gamma)\, p_t(a)}. \tag{6.4}$$

4. Contextual Bandit

Regret Analysis

- **Regret:** $\mathbb{E}[R(T)] \leq O(\sqrt{KT \log N})$
- **Proof:**

Each realization of $\widehat{G}_t$ satisfies:

$$
\begin{aligned}
\widehat{G}_t &:= \sum_{e \in \mathcal{E}} p_t(e)\, \widehat{c}_t^2(e) \\
&= \sum_a \sum_{e \in \mathcal{E}_a} p_t(e) \cdot \widehat{c}_t(e) \cdot \widehat{c}_t(e) && \textit{(re-write as a sum over arms)} \\
&\leq \sum_a \sum_{e \in \mathcal{E}_a} \frac{p_t(e)}{(1-\gamma)\, p_t(a)}\, \widehat{c}_t(a) && \textit{(replace one } \widehat{c}_t(a) \textit{ with an upper bound (6.4))} \\
&= \frac{1}{1-\gamma} \sum_a \frac{\widehat{c}_t(a)}{p_t(a)} \sum_{e \in \mathcal{E}_a} p_t(e) && \textit{(move “constant terms” out of the inner sum)} \\
&= \frac{1}{1-\gamma} \sum_a \widehat{c}_t(a) && \textit{(the inner sum is just } p_t(a))
\end{aligned}
$$

To complete the proof, take expectations over both sides and recall that $\mathbb{E}[\widehat{c}_t(a)] = c_t(a) \leq 1$. □

4. Contextual Bandit

Regret Analysis

- **Regret:** $\mathbb{E}[R(T)] \leq O(\sqrt{KT \log N})$
- **Proof:**

Let us complete the analysis, being slightly careful with the multiplicative constant in the regret bound:

$$
\begin{aligned}
\mathbb{E}\left[\widehat{R}_{\texttt{Hedge}}(T)\right] &\leq 2\sqrt{3/(1-\gamma)} \cdot \sqrt{TK \log N} \\
\mathbb{E}\left[R_{\texttt{Exp4}}(T)\right] &\leq 2\sqrt{3/(1-\gamma)} \cdot \sqrt{TK \log N} + \gamma T && \textit{(by Eq. (6.3))} \\
&\leq 2\sqrt{3} \cdot \sqrt{TK \log N} + 2\gamma T && \textit{(since } \sqrt{1/(1-\gamma)} \leq 1+\gamma\textit{)} \qquad (6.5)
\end{aligned}
$$

4. Contextual Bandit

Motivation

- **Motivation:** The expert is the context so far!
- **Examples**: "User profile", feature of the environment (day of the week, season, proximity to a major event), features of their own, indicate the set of feasible features.

4. Contextual Bandit

Notation

- x_t: **context**, $\mathscr{X}$: **set of contexts**
- a_t: **action given the context**, $\mathscr{A}$: **set of actions**, $|\mathscr{A}| = K$
- l_t: **loss**
- $\pi : \mathscr{X} \to \mathscr{A}$: **a policy**, Π: **a class of policy**
- **Regret**: $R(T) = \sum_{t=1}^{T} l_t(a_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} l_t(\pi(x_t))$
  - **Stochastic bandit**: $l_t \sim \mathscr{P}(. \,|\, x_t, a_t)$

4. Contextual Bandit

Revisted

- **Regret:** $\mathbb{E}[R(T)] \leq O(\sqrt{KT \log N})$
- **Drawbacks**: If $|\Pi|$ is continuous, then huge regret! Ignore the structure of the context!

4. Contextual Bandit

Notation

- θ_a^*: **coefficient vector** ($||\theta_a^*||_2 \leq S$)
- $x_{a,t}$: **context for each arm**
- $l_{a,t} = x_{a,t}^T \theta_a^* + \epsilon_t$ with $\epsilon_t \sim$ R-sub-Gaussian
- $X_{a,t} = \begin{bmatrix} x_{a,1}^T \\ \cdots \\ x_{a,m}^T \end{bmatrix}$ (Suppose pull m times before), $\Gamma_{a,t} = \begin{bmatrix} l_{a,1} \\ \cdots \\ l_{a,m} \end{bmatrix}$
- $b_{a,t} = X_{a,t}^T \Gamma_{a,t}$

4. Contextual Bandit

Notation

- **Ridge regression estimator**: $\hat{\theta}_{a,t} = (X_{a,t}^T X_a + \lambda I)^{-1} b_{a,t}$
- $A_{a,t} = I\lambda + \sum_{s=1}^{t} x_{a,s} x_{a,s}^T$ (λ - regularization parameter related to sub-Gaussian variable)
- $\theta_a^* \in C_{a,t}$ where

  $$C_{a,t} = \{\theta_a^* \in \mathbb{R}^d : ||\hat{\theta}_{a,t} - \theta_a^*||_{A_{a,t}} \leq R\sqrt{2\log(\frac{\det(A_{a,t})^{1/2}\det(\lambda I)^{-1/2}}{\delta})} + \lambda^{1/2} S\}$$

- $a_t = \arg\max_a \max_{\theta_a \in C_{a,t-1}} x_{a,t}^T \theta_a$

4. Contextual Bandit

Algorithm

Algorithm 3 LinUCB with Contextual Bandits

Input: $R \in \mathbb{R}^+$, regularization parameter λ

for $t = 1, 2, \ldots, T$ **do**

  Observe feature vectors of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{a,t} \in \mathbb{R}^d$

  for all $a \in \mathcal{A}_t$ **do**

    if a is new **then**

      $\mathbf{A}_a \leftarrow \lambda \mathbf{I}_d$ (d-dimensional identity matrix)

      $\mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1}$ (d-dimensional zero vector)

    end if

    $\hat{\theta}_a \leftarrow \mathbf{A}_a^{-1}\mathbf{b}_a$

    $C_{a,t} \leftarrow \left\{ \theta_a^* \in \mathbb{R}^d : \left|\left|\hat{\theta}_{a,t} - \theta_a^*\right|\right|_{\mathbf{A}_a} \leq R\sqrt{2\log\left(\frac{\det(\mathbf{A}_a)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2}S \right\}$

    $p_{a,t} \leftarrow \arg\max_{\hat{\theta}_a \in C_{a,t}} \mathbf{x}_{a,t}^T \hat{\theta}_a$

  end for

  Choose arm $a_t = \arg\max_{a\in\mathcal{A}_t} p_{a,t}$ with ties broken arbitrarily, and observe payoff r_t

  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{a_t,t}\mathbf{x}_{a_t,t}^T$

  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{a_t,t}$

end for

4. Contextual Bandit

Regret analysis

- $R_n = O(RS\lambda^{1/2}dK \ln(T/\delta)\sqrt{T})$
- **Proof**: Skipped due to martingale properties!

5. Bandit with Knapsacks

Motivation

- **Dynamic pricing**:
    - **A seller**: B copies of **an item**.
    - **A customer**: Buy items
    - For round $t = 1,..,T$
        - A seller put a price p_t on an item, while the customer values v_t on that item.
        - If $v_t \geq p_t$, then the customer will buy the item. Otherwise, there is no sale.
        - If the customer buy the item, the seller will receive p_t as a reward.
        - The algorithm stop when either $t = T$ or no more copies of the item.
        - **Objective**: $\max \sum_{t=1}^{T} p_t$
        - **Remark**: B < T

5. Bandit with Knapsacks

Formalization

Problem protocol: Bandits with Knapsacks (BwK)

Parameters: K arms, d resources with respective budgets $B_1, \ldots, B_d \in [0, T]$.
In each round $t = 1, 2, 3 \ldots$:

1. Algorithm chooses an arm $a_t \in [K]$.
2. Outcome vector $\vec{o}_t = (r_t; c_{t,1}, \ldots, c_{t,d}) \in [0,1]^{d+1}$ is observed,
   where r_t is the algorithm's reward, and $c_{t,i}$ is consumption of each resource i.

Algorithm stops when the total consumption of some resource i exceeds its budget B_i.

- **Remark**: Outcome vector given a selected arm is sampled i.i.d.
- **Remark:** Time is also a resource.
- **Goal**: Maximize the total reward.

5. Bandit with Knapsacks

Formalization - Different from regular bandit setting

- **Example (Motivation)**: 2 arms, 2 products, each has M budgets, each arm will select one of the product, consume 1 unit and get 1 reward. Assume T > 2M.
    - For regular bandit, we can **choose either arm** and play repeatedly! The total rewards will be **M**.
    - However, **alternating** two arms can lead to **2M rewards**!
- **Key difference**:
    - Non-systematic exploration at the beginning can be harmful in the long-term due to constraint budget.
    - Expected per-round reward is no longer a right objective because of high resource consumption.
    - Fixed arm is not longer suitable, rather a distribution of arms.

5. Bandit with Knapsacks

Examples

- **Dynamic pricing**: Suppose $d = 1$ with resource is B_1. K Arms corresponds to K different prices. Goal is to maximize $\sum_{t=1}^{T} p_t$
  - $o_t = \begin{cases} (p_t, 1) & p_t \geq v_t \\ (0,0) & \text{Otherwise} \end{cases}$
- **Dynamic pricing for hiring**: Suppose $d = 1$ job with budget B_1. K arms corresponds to K different prices. Goal is to maximize $\sum_{t=1}^{T} 1[p_t \geq v_t]$
  - $o_T = \begin{cases} (1, p_t) & p_t \geq v_t \\ (0,0) & \text{otherwise} \end{cases}$
- **Pay-per-click ad allocation**: Suppose $d = 1$ site with budge B_1. K arms corresponds to K ads. Goal is to maximize the number of clicks
  - $o_t = \begin{cases} (r_a, r_a) & \text{if click with probability } q_a \\ (0,0) & \text{Otherwise} \end{cases}$

5. Bandit with Knapsacks

Algorithms - Primal-Dual methods

- **Linear relaxation**: Consider a fixed distribution D and outcomes o equals to expected value of outcomes. Then, optimizing D using linear programming.
- $r(D) = \sum_a D(a)r(a), c_i(D) = \sum_a D(a)c_i(a)$
- $\max r(D)$ subject to $D \in \Delta_K, T \cdot c_i(d) \leq B \forall i \in [d]$
- **Corollary**: $T \cdot OPT_{LP} \geq OPT$

5. Bandit with Knapsacks

Algorithms - Primal-Dual methods

- **Langrange game:**
  - **Lagrange function:** $L(D, \lambda) = r(D) + \sum_{i \in D} \lambda_i (1 - \frac{T}{B} c_i(D))$ where $\lambda \in \Delta^d$
  - **Langrange game**: zero-sum game, primal player selects action, dual player selects resources, and the payoff $L(a, i) = r(a) + 1 - \frac{T}{B} c_i(a)$
  - **Lemma:**
    - $D*$ is optimal for the linear programming.
    - $1 - \frac{T}{B} c_i(D*) \geq 0$, with the equality if $\lambda_i^* > 0$
    - $L(D*, \lambda*) = OTP_{LP}$

5. Bandit with Knapsacks

Algorithms - Primal-Dual methods

- **Repeated Lagrange game**: $L_t(a, i) = r_t(a) + 1 - \frac{T}{B} c_{t,i}(a)$
  - **Remark**: $\mathbb{E}[L_t(a, i)] = L(a, i)$

Given : time horizon T, budget B, number of arms K, number of resources d.
Bandit algorithm $\texttt{ALG}_1$: action set $[K]$, maximizes rewards, bandit feedback.
Bandit algorithm $\texttt{ALG}_2$: action set $[d]$, minimizes costs, full feedback.
for *round* $t = 1, 2, \ldots$ *(until stopping)* **do**
$\texttt{ALG}_1$ returns arm $a_t \in [K]$, algorithm $\texttt{ALG}_2$ returns resource $i_t \in [d]$.
Arm a_t is chosen, outcome vector $\vec{o}_t = (r_t(a_t); c_{t,1}(a_t), \ldots, c_{t,d}(a_t)) \in [0, 1]^{d+1}$ is observed.
The payoff $\mathcal{L}_t(a_t, i_t)$ from (10.10) is reported to $\texttt{ALG}_1$ as reward, and to $\texttt{ALG}_2$ as cost.
The payoff $\mathcal{L}_t(a_t, i)$ is reported to $\texttt{ALG}_2$ for each resource $i \in [d]$.
end

Algorithm 10.1: Algorithm $\texttt{LagrangeBwK}$

5. Bandit with Knapsacks

Algorithms - Primal-Dual methods

- **Regret:** Suppose ALG_1 is $EXP3$ and ALG_2 is $Hedge$, then the regret bound is achieved with the probability at least $1 - \delta$

$$OPT - REW \leq O(T/B)\sqrt{TK \ln \frac{dT}{\delta}}$$

- Optimal **only** in the region $B = \Omega(T)$

5. Bandit with Knapsacks

Algorithms - UCB-like methods

- Let $M_t = \begin{bmatrix} r_t(a_1) & c_{1,t}(a_1) & \dots & c_{d,t}(a_1) \\ \cdots & \cdots & \cdots & \cdots \\ r_t(a_k) & c_{t,1}(a_k) & \dots & c_{t,d}(a_k) \end{bmatrix}$
- Define ConfReg is a confidence region around M_t,
- $UCB_t(D \mid B) = \sup_D LP(D \mid B, M)$ for some $M \in$ ConfReg.

Rescale the budget: $B' \leftarrow B(1-\epsilon)$, where $\epsilon = \tilde{\Theta}(\sqrt{K/B})$
Initialization: pull each arm once.
for *all subsequent rounds t* **do**
    In each round t, pick distribution $D = D_t$ with highest $\texttt{UCB}_t(\cdot \mid B')$.
    Pick arm $a_t \sim D_t$.
end

Algorithm 10.3: `UcbBwK`: Optimism under Uncertainty with Knapsacks.

5. Bandit with Knapsacks

Algorithms - UCB-like methods

- **Regret**: $OPT - \mathbb{E}[REW] \leq O(\sqrt{K \cdot OPT} + OPT\sqrt{\frac{K}{B}})$
  - **Remark**: $\sqrt{K \cdot OPT}$ is similar to stochastic bandit!
  - **Remark**: Optimal for any given triple (K, B, T)

5. Bandit with Knapsacks

Generalization - Contextual bandits

- **LagrangeBwK**:
    - $OPT_\Pi - \mathbb{E}[REW] \le \tilde{O}(T/B)\sqrt{KT \log |\Pi|}$
    - ALG1 is Exp4, but not computationally efficient.
- **Successive elimination and Policy elimination**:
    - $OPT_\Pi - \mathbb{E}[REW] \le \tilde{O}(1 + OPT_\Pi/B)\sqrt{KT \log |\Pi|}$
    - Not computationally efficient.
- **UcbBwK** (Linear contextual bandit):
    - $OPT - \mathbb{E}[REW] \le \tilde{O}(m\sqrt{T})(1 + OPT/B)$ in regime $B > mT^{3/4}$
    - Not computationally efficient.

5. Bandit with Knapsacks

Generalization - Bandit convex optimization

- $o_t = (f_t, g_{t,1}, \ldots, g_{t,d})$ where $f_t : \mathcal{X} \to [0,1], g_{t,i} : \mathcal{X} \to [0,1]$ and $\mathcal{X} \subset \mathbb{R}^K$
  - f_t is a concave function, $g_{t,i}$ is a convex function, $\mathcal{X}$ is a convex set.
  - **LagrangeBwK**: $\frac{T}{B}\sqrt{T} \cdot poly(K \log T)$

5. Bandit with Knapsacks

Generalization - Adversarial

- Sequence $M_1, \ldots, M_T$ are fixed, not sampling, before round 1.
- **Challenge**: How much budget to save for the future, without being able to predict.
- **Competitive ratio**: $\dfrac{OPT_{FD}}{\mathbb{E}[REW]}$
- **Modified LagrangeBwK**: $(OPT_{FD} - reg)/\mathbb{E}[REW] \le O_d(\log T)$
    - $reg = O(1 + \dfrac{OPT_{FD}}{dB})\sqrt{TK \log(Td)}$
    - **Remark:** Time is not include in the outcome matrices
    - **Remark:** $\dfrac{T}{B}$ is replaced by $\gamma \in (0, \dfrac{T}{B}]$, which is sampled at random

5. Bandit with Knapsacks

Generalization - Best in both worlds

- **Corrupted environments**: mixed between adversarial and stochastic environment!
- Recent researches seem to focus on this problem (Neurips, ICML, ICLR)