1-Bandits with Delayed Feedback

Jingqi Fan

November 13, 2023

1. Setting
   - Setting about Bandits
   - Setting about Delay
2. Algorithm
3. Technique

- stochastic
- adversarial
- linear
- two-player
- best-of-both-worlds
- combinatorial semi-bandit

- best-of-both-worlds
  1. A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback (2022)
  2. An Improved Best-of-both-worlds Algorithm for Bandits with Delayed Feedback (2023)

A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback

Prelimilaries:

- at time $t = 1, 2, \ldots$, the learner chooses an arm I_t among K arms
- it suffers a loss ℓ_{t,I_t} from a loss vector $\ell_t \in [0, 1]^K$ generated by the environment but does not observe it
- After d_t, the learner observed the pair (t, ℓ_{t,I_t}) at the end of round $t + d_t$
- w.l.o.p. assume $t + d_t \leq T$ for all t
- consider two regimes, oblivious adversarial and stochastic

A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback

Pseudo-regret

$$\bar{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T}(\ell_{t,I_t} - \ell_{t,i_T^*})\right] \tag{1}$$

where $i_T^* \in \arg\min_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$

Algorithm(FTRL):

Algorithm 1: FTRL with advance tuning for delayed bandit

Input: Learning rate rule η_t and γ_t
Initialize $\mathcal{D}_0 = 0$ *and* $\hat{L}_1^{obs} = \mathbf{0}_K$ *(where* $\mathbf{0}_K$ *is a zero vector in* $\mathbb{R}^K$*)*
for $t = 1, \ldots, n$ **do**
  determine γ_t
    Set $\sigma_t = \sum_{s=1}^{t-1} \mathbb{1}(s + d_s > t)$
    Update $\mathcal{D}_t = \mathcal{D}_{t-1} + \sigma_t$
  Set $x_t = \arg\min_{x \in \Delta^{K-1}} \langle \hat{L}_t^{obs}, x \rangle + F_t(x)$
  Sample $I_t \sim x_t$
  for $s : s + d_s = t$ **do**
    Observe (s, ℓ_{s,I_s})
    Construct $\hat{\ell}_s$ and update $\hat{L}_t^{obs}$

- $\sigma_t = \sum_{s=1}^{t-1} \mathbb{I}(s + d_s \geq t)$ is the number of outstanding observations and $\mathcal{D}_t = \sum_{s=1}^{t} \sigma_t$
- we set $\eta_0 = 10 d_{\max} + \frac{d_{\max}^2}{(K^{1/3} \log(K))^2}$ and $\gamma_0 = 24^2 d_{\max}^2 K^{2/3} \log(K)$
- similar hybrid regularizer $F_t(x)$ as Zimmert and Seldin (2020)

A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback

Regret Bound:

Setting	Assumption	Regret Bound
stochastic	fixed delay	$O\left(\sum_{i\neq i^*}(\frac{\log T}{\Delta_i}+\frac{d}{\Delta_i \log K})+dK^{1/3}\log K\right)$
adversarial	fixed delay	$O\left(\sqrt{KT}+\sqrt{dT\log K}\right)$
stochastic	arbitrary delay with known $d_{\max}$	$O\left(\sum_{i\neq i^*}(\frac{\log T}{\Delta_i}+\frac{\sigma_{\max}}{\Delta_i \log K})+d_{\max}K^{1/3}\log K\right)$
adversarial	arbitrary delay with known $d_{\max}$	$O\left(\sqrt{KT}+\sqrt{D\log K}+d_{\max}K^{1/3}\log K\right)$

An Improved Best-of-both-worlds Algorithm for Bandits with Delayed Feedback

- similar setting and algorithm with the former paper
- remove the assumption of known $d_{\max}$ by using skipping trick to leave out too large delay

- combinatorial semi-bantit
  1. Non-stationary Delayed Combinatorial Semi-Bandit with Causally Related Rewards (2023)
  2. A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs (2023)

A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs

Prelimilaris:

- In each round $t \in [T]$ the learner chooses an action $\boldsymbol{a}_t \in \mathcal{A} \subseteq \{0,1\}^K$
- it suffers loss $\boldsymbol{a}_t^\top \boldsymbol{\ell}_t$, where $\boldsymbol{\ell}_t \in \mathbb{R}^K$
- then the learner observes $\{\mathcal{L}(\boldsymbol{\ell}_\tau, \boldsymbol{a}_\tau) : \tau + d_\tau = t\}$, defined as $\mathcal{L}(\boldsymbol{\ell}_\tau, \boldsymbol{a}_\tau) = \boldsymbol{a}_\tau \odot \boldsymbol{\ell}_\tau$
- delay $d_1, ..., d_T$ and loss $\ell_1, ..., \ell_T$ ar both generated by an oblivious adversary
- $o_t = \{\tau : \tau + d_\tau < t\}$ is the set of indices of observed losses at $t-1$ and $m_t = [t-1] \backslash o_t$ not be observed

A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs

Assumption:

- $d_{\max} = \max_{t \in [T]} d_t \geq 1$ is known to the learner
- $\sum_{t=1}^{T} |m_t| = D$ and T is known to the learner

pseudo-regret

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} (\boldsymbol{a}_t - \boldsymbol{a}^*)^\top \boldsymbol{\ell}_t\right] \tag{2}$$

A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs

Algorithm(FTRL):

$$\boldsymbol{w}_t = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \sum_{\tau \in \boldsymbol{o}_t} \hat{\boldsymbol{\ell}}_\tau^\top \boldsymbol{w} + R(\boldsymbol{w}) \tag{3}$$

where $o_t = \{\tau : \tau + d_\tau < t\}$ is observed index set and

$$R(\boldsymbol{w}) = \sum_{i=1}^{K} \left(\frac{1}{\eta} \boldsymbol{w}(i) \log(\boldsymbol{w}(i)) - \frac{1}{\eta} \log(\boldsymbol{w}(i)) \right) \tag{4}$$

Regret Bound:

$$O(\sqrt{B(KT + BD)\log(K)}) \tag{5}$$

where $D = \sum_{t=1}^{T} d_t$ and $\max_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_1 \leq B$

A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs

Handling unkonwn $d_{\max}$: Double Trick

- $\mathcal{T}_e = \{t : 2^{e-1} \leq \max_{j \in o_t} d_j \leq 2^e\}$ is the set of indices of epoch e
- $\tilde{\mathcal{T}}_e = \{t \in \mathcal{T}_e : d_t \leq 2^e\}$ is the indices of epoch e with delay $\leq 2^e$
- Then $R_{T,D}(2^e) \leq R_{T,D}(2d_{\max})$
- the regret in $\mathcal{T}_e \backslash \tilde{\mathcal{T}}_e$ is at most $Md_{\max}$ since $|\mathcal{T}\backslash \tilde{\mathcal{T}}_e| \leq d_{\max}$

see Bistritz et al.(2019) handling unknown T and D

- fixed delay
- $d_{\max}$ known, T known (for parameter)
- unbounded delay
- composite anonymous delay
- arm-dependent delay
- reward-dependent delay
- instant reward combined with delay

Unbounded delay $\rightarrow$ count outstanding delay

1. doubling trick: Online EXP3 Learning in Adversarial Bandits with Delayed Feedback (2019)
2. skipping technique: An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays (2020)

Online EXP3 Learning in Adversarial Bandits with Delayed Feedback

Prelimilaries:

- at each round t a player picks one out of K arms, denoted as a_t
- the cost at round t from arm i is $l_t^{(i)} \in [0, 1]$ and $\boldsymbol{l}_t = \left(l_t^{(1)}, ..., l_t^{(K)}\right)$ is the cost vector
- Cost and delay are both chosen adversarially
- $\mathcal{S}_t$ is the set of costs received at round t
- $\mathcal{M}_t$ is the set of missing samples s.t. $t + d_t > T$
- the vector of probabilities of the player for choosing arms at t is $\boldsymbol{p}_t \in \Delta^K$

Online EXP3 Learning in Adversarial Bandits with Delayed Feedback

Algorithm 1 EXP3 with delays

Initialization: Let $\{\eta_t\}$ be a positive non-increasing sequence, and set $\tilde{L}_1^{(i)} = 0$ and $p_1^{(i)} = \frac{1}{K}$ for $i = 1, ..., K$.

For $t = 1, ..., T$ **do**

1. Choose an arm a_t at random according to the distribution $\boldsymbol{p}_t$.
2. Obtain a set of delayed costs $l_s^{(a_s)}$ for all $s \in \mathcal{S}_t$, where a_s is the arm played at round s.
3. Update the weights of arm a_s for all $s \in \mathcal{S}_t$, using
$$\tilde{L}_t^{(a_s)} = \tilde{L}_{t-1}^{(a_s)} + \eta_s \frac{l_s^{(a_s)}}{p_s^{(a_s)}}. \quad (3)$$
4. Update the mixed strategy
$$p_{t+1}^{(i)} = \frac{e^{-\tilde{L}_t^{(i)}}}{\sum_{j=1}^{n} e^{-\tilde{L}_t^{(j)}}}. \quad (4)$$

End

Online EXP3 Learning in Adversarial Bandits with Delayed Feedback

Use double trick to handle unknown D and T

- idea: start a new epoch every time $\sum_{\tau}^{t} m_\tau$ doubles, where m_t is the number of missing feedback samples at t
- define the e-epoch as

$$\mathcal{T}_e = \left\{ t | 2^{e-1} \leq \sum_{\tau=1}^{t} m_\tau \leq 2^e \right\} \tag{6}$$

- Then the sum of delays is within a given interval and in every epoch e, set $\eta_e = \sqrt{\frac{\ln K}{2^e}}$ to get adaptive algorithm

Online EXP3 Learning in Adversarial Bandits with Delayed Feedback

Analysis about Double Trick:

- Define $\mathcal{M}_e$ as the set of feedback for costs in epoch e that are not received within epoch e
- $T_e = \max \mathcal{T}_e$ is the last round in $\mathcal{T}_e$

$$\sum_{t \in \mathcal{T}_e, t \notin \mathcal{M}_e} d_t \leq \sum_{\tau = T_{e-1}+1}^{T_e} m_\tau \leq 2^{e-1} \tag{7}$$

- Then apply the regret of algorithm1 separately on every epoch and yield:

$$R_e \triangleq E^{\boldsymbol{a}} \left\{ \sum_{t \in \mathcal{T}_e} l_t^{(a_t)} - \min_i \sum_{t \in \mathcal{T}_e} l_t^{(i)} \right\} \leq \frac{\ln K}{\eta_e} + \eta_e \left(\frac{K}{2} |\mathcal{T}_e| + 4 \sum_{t \in \mathcal{T}_e, t \notin \mathcal{M}_e} d_t \right) + |\mathcal{M}_e|$$

Online EXP3 Learning in Adversarial Bandits with Delayed Feedback

- the "cheapest" way to increase $|\mathcal{M}_e|$ is when the feedback from T_e is delayed by 1, from $T_e - 1$ delayed by 2 and so on

$$\sum_{i=1}^{|\mathcal{M}_e} i = \frac{|\mathcal{M}_e(|\mathcal{M}_e| + 1)|}{2} \leq 2^{e-1} \tag{8}$$

Finally the regret bound is:

$$O\left(\sqrt{\ln K \left(KT + \sum_{t=1}^{T} d_t\right)}\right) \rightarrow O\left(\sqrt{\ln K \left(K^2 T + \sum_{t=1}^{T} d_t\right)}\right) \tag{9}$$

An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays

Prelimilaries:

- At time $t = 1, ..., n$ the learner picks $A_T \in [k]$
- it immediately suffers ℓ_{t,A_t}, where $(\ell_t)_{t=1,...,n}$ are vectors in $[0,1]^k$.
- environment chooses a sequence of delay $(d_t)_{t=1,...,n}$
- the player observes the tuples (s, ℓ_{s,A_s}) for each s s.t. $s + d_s = t$ at the end of time t.
- w.l.o.p. assume $t + d_t \leq n$ for all t
- focus on oblivious adversarial setting on reward and delay

An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays

Regret:

$$\mathcal{R}_n := \mathbb{E}\left[\sum_{t=1}^{n} \ell_{t,A_t} - \min_{i\in[k]} \sum_{t=1}^{n} \ell_{t,i}\right] \tag{10}$$

Definitions:

- $D = \sum_{t=1}^{n} d_t$ is the total delay
- for a set $S \subset [n] = \{1, ..., n\}$, its complement is $\bar{S} = [n]\backslash S$
- for a convex function F, F^* denotes its convex conjugate, and $\bar{F}^*$ is constrained convex conjugate, defined as

$$F^*(y) = \max_{x\in\mathbb{R}^k} < x, y > -F(x) \tag{11}$$

$$\bar{F}^*(y) = \max_{x\in\Delta([k])} < x, y > -F(x) \tag{12}$$

An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays

Algorithm(FTRL):

- hybrid regularizers:

$$F_t(x) = -\sum_{i=1}^{k} 2\sqrt{t} x_i^{1/2} + \eta_t^{-1} \sum_{i=1}^{k} x_i \log(x_i) \tag{13}$$

1. $\frac{1}{2}$-Tsallis Entropy + negative entropy
2. same decomposition as $\Omega(\max\{\sqrt{kn}, \sqrt{dn\log(k)}\})$ (Cesa-Bianchi et al.)
3. future tune the learning rate η by skipping trick

An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays

Algorithm(FTRL):

- count outstanding delay(tuning learning rate to unknown D setting):
  1. simple tuning
  2. advanced tuning (skipping)

Simple Tuning

Algorithm 1: FTRL for bandits with delay

Input: Proper learning rate rule η_t
Initialize $\hat{L}_1^{obs} = 0$
Initialize $\mathfrak{D}_0 = 0$ (simple tuning)
for $t = 1, \dots, n$ **do**
  determine η_t
    Set $\mathfrak{D}_t = \mathfrak{D}_{t-1} + \mathfrak{d}_t$
    Set $\eta_t^{-1} = \sqrt{2\mathfrak{D}_t / \log(k)}$ (simple tuning)
  Set $x_t = \arg\min_{x \in \Delta([k])} \langle x, \hat{L}_t^{obs} \rangle + F_t(x)$
  Sample $A_t \sim x_t$
  for $s : s + d_s = t$ **do**
    Observe (s, ℓ_{s,A_s})
    Construct $\hat{\ell}_s$ and update $\hat{L}_t^{obs}$

- $\mathfrak{d}_t = \sum_{s=1}^{t-1} \mathbb{I}\{s + d_s \geq t\}$ is the number of outstanding observations at round t
- $\mathfrak{D}_t = \sum_{s=1}^{t} \mathfrak{d}_t$ and $\eta_t^{-1} = \sqrt{\frac{2\mathfrak{D}_t}{\log(k)}}$
- well-defined when $\mathfrak{D}_t = 0$

Advanced Tuning (skipping)

Algorithm 2: Advanced tuning of η_t for Alg. 1

Initialize $\tilde{\mathfrak{D}}_0 = 0$ and $(a_s^t)_{s=1,\dots,n;t=1,\dots,n} = 1$
determine η_t
  Set $\tilde{\mathfrak{d}}_t = \sum_{s=1}^{t-1} \mathbb{I}\{s + d_s \geq t\} a_s^t$
  Update $\tilde{\mathfrak{D}}_t = \tilde{\mathfrak{D}}_{t-1} + \tilde{\mathfrak{d}}_t$
  Set $\eta_t^{-1} = \sqrt{\tilde{\mathfrak{D}}_t / \log(k)}$
  for $s = 1, \dots, t-1$ **do**
    if $\min\{d_s, t-s\} > \eta_t^{-1}$ **then**
      $(a_s^{t'})_{t'>t} = 0$ (At most one index s satisfies the **if**-condition, see Lemma 5)

- the optimal subset of rounds $\bar{S}$ and the remaining rounds $S = [n] \backslash \bar{S}$
- modify $\mathfrak{d}_t$ by skipping some outstanding observations but still using them to estimate $\mathfrak{D}_t$
- Hence we define indicator variables $a_s^t \in \{0, 1\}$
- $\tilde{\mathfrak{d}}_t = \sum_{s=1}^{t-1} a_s^t \mathbb{I}\{s + d_s \geq t\}$

Advanced Tuning (skipping)

Intuition behind the skipping precedure:

- Cesa-Bianchi et al. (2016) has $O(\sqrt{kn} + \sqrt{D\log(k)})$
- Then $O(\sqrt{kn} + |S| + \sqrt{D_{\bar{S}}\log(k)})$ in delay setting
- $|S| = \sqrt{D_{\bar{S}}\log(k)}$ is the number of skipped rounds
- $D_S = \sum_{t\in S} d_t \geq X|S| \geq D_{\bar{S}}$ and we get threshold X:

$$X \geq \sqrt{\frac{D_{\bar{S}}}{\log(k)}} \tag{14}$$

- finally we replace $D_{\bar{S}}$ with $\tilde{\mathfrak{D}}_t$

An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays

Regret Bound of Algorithm1

$$O(\sqrt{kn} + \sqrt{D\log(k)}) \tag{15}$$

Regret Bound of Algorithm2

$$O(\sqrt{kn} + \min_{S}(|S| + \sqrt{D_{\bar{S}}\log(k)})) \tag{16}$$

match the lower bound in Cesa-Bianchi et al.(2016) and refine it

Composite Anonymous Delay:

1. Nonstochastic Bandits with Composite Anonymous Feedback (2018)
    - oblivious setting
2. Bandits with Delayed, Aggregated Anonymous Feedback (2018)
    - stochastic setting
3. Adaptive Algorithms for Multi-armed Bandit with Composite and Anonymous Feedback (2020)
    - stochastic setting
    - non-obivious setting
4. Bounded Memory Adversarial Bandits with Composite Anonymous Delayed Feedback (2022)
    - oblivious setting

Stochastic Composite Anonymous Delay

Setting:

- the loss observed at the end of each round is a sum of many loss components of previous actions
- In each time a player chooses one arm among $\mathcal{N} = \{1, ..., N\}$.
- the arm i generates an i.i.d. reward vector in $\mathbb{R}_+^\infty$
- $\boldsymbol{r}_{a(t)}(t) = (r_{(t),1}(t), r_{(t),2}(t), ...)$, where $\boldsymbol{r}_{a(t),\tau}(t)$ is the partial reward
- $D_{a(t)}$ is the distribution of $\boldsymbol{r}_{a(t)}$ and define $\boldsymbol{\mu}_{a(t)} := \mathbb{E}_{D_{a(t)}}[\boldsymbol{r}_{a(t)}]$ as its mean
- Then at every time the player receove the aggregated reward from previous arms, i.e. $Y(t) := \sum_{\tau \leq t-1} r_{a(\tau),t-\tau}(\tau)$

Stochastic MAB with Composite Anonymous Delay

- $s_i := \|\boldsymbol{\mu}_i\|_1$ is the expected total reward of pulling arm i
- w.l.o.p. assume $1 \geq s_1 \geq s_s \geq ... \geq s_N \geq 0$
- denote $\Delta_i := s_1 - s_i$ for all $i \geq 2$ is the reward gap of arm i

The Cumulative Regret

$$Reg(T) := Ts_1 - \mathbb{E}\left[\sum_{t=1}^{T} s_{a(t)}\right] \tag{17}$$

Non-oblivious Adversarial MAB with Composite Anonymous Delay

Setting:

- $\mathcal{N} = \{1, ..., N\}$ is the arm set and $\|\boldsymbol{r}_i(t)\|_1 \leq 1$
- non-oblivious delay: the actual reward is oblivious but the spread of reward is non-oblivious as long as $\|\boldsymbol{r}_i(t)\|_1 = s_i(t)$
- aggregated reward $Z(t) := \sum_{\tau=t-d}^{t-1} r_{a(\tau),t-\tau}(\tau)$
- $G_i := \sum_{t=1}^{T} \|\boldsymbol{r}_i(t)\|_1$

The total regret:

$$Reg(T) := \mathbb{E}\left[\max_i G_i\right] - \mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{r}_{a(t)}(t)\|_1\right] \tag{18}$$

- arm-dependent delay
  1. Stochastic bandits with arm-dependent delays (2020) (also heavy-tailed delays)
  2. Nonstochastic Bandits and Experts with Arm-Dependent Delays (2021)
- reward-dependent delay
  1. Stochastic Multi-Armed Bandits with Unrestricted Delay Distributions (2021)
- both reward and delay
  1. A New Framework: Short-Term and Long-Term Returns in Stochastic Multi-Armed Bandit (2023)

Arm-dependent Delay

Nonstochastic Setting:

- partially-concealed bandit: At the end of round $t + d_t(i_t)$, after pull a arm i_t, the learner receives
  1. loss $\ell_t(i_t)$
  2. the number of missing observations $\rho_t(i_t) = |\{s : s < t, s + d_s(i_t) \geq t\}|$
- concealed bandit: the learner just receives $\ell_t(i_t)$

Expected Regret

$$\mathbb{E}[\mathcal{R}_T(\boldsymbol{u})] = \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(i_t) - < \boldsymbol{u}, \boldsymbol{\ell}_t >)\right] \tag{19}$$

Arm-dependent Delay

Why partially-concealed or concealed:

- mild assumptions than other work

Arm-dependent Delay

Stochastic Setting:

- finite arm set $K \in \mathbb{N}^*$ and $[K] \triangleq \{1, ..., K\}$
- each arm $i \in [K]$ is associated with both
  1. an unknown reward distribution $\mathcal{V}_i$ in [0,1] with mean μ_i
  2. an unknown delay distribution $\mathcal{D}_i$ with cumulative distribution function τ_i supported in $\mathbb{N}$ s.t. for any $d \geq 0, t \leq T$, if $D_t \sim \mathcal{D}_i$, then $\mathbb{P}(D_t \leq d) = \tau_i(d)$
- $C_t \sim \mathcal{V}_{I_t}, D_t \sim \mathcal{D}_{I_t}$
- at round $t + u$ for $1 \leq u \leq T - t$, the learner only observes

$$X_{t,u} \triangleq C_t \mathbb{I}\{D_t \leq u\} \tag{20}$$

Arm-dependent Delay

Challenge:

- the ambiguity on either $C_s = 0$ or $t - s < D_s$
- i.e. the learner cannot know exactly how much feedback is missing
- after several rounds, the actual reward is scaled:

$$\mathbb{E}\left[X_{u,t-u} | I_u = i\right] = \tau_i(t-u)\mu_i \tag{21}$$

Heavy-tailed Delay

Assumption 1

Let $\alpha > 0$ be some fixed quantity, we assumpe that $\forall m \in \mathbb{N}^*$ and $\forall i \in \{1, ..., K\}$, it holds that

$$|1 - \tau_i(m)| \leq m_{-\alpha} \tag{22}$$

the smaller α, the more heavy-tailed the delay distribution

Expected Regret

$$\bar{R}_T = T\mu^* - \mathbb{E}\sum_{t=1}^{T} C_t = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_i(T)\right] \tag{23}$$

Reward-dependent Delay

- there is no restriction on the joint distribution between $r_t(\cdot)$ and $d_t(\cdot)$ (i.e. remove the assumption of independence)
- Main challenge: the observed emprical mean is no long an unbiased exstimator

Instant Reward Combined with Delay

Setting: The distributions of reward for short-term and longterm rewards are related with a previously **known** relationship. (linear in this paper)

- At each round t, the observer pulls an arm $a \in \{1, ..., K\}$
- it observes an instant reward $f_t(a_t)$ and generates a delayed reward $r_t(a_t)$, which will be observed after $d_t(a_t)$ rounds.
- $r_t(i) \sim R_i$, $f_t(i) \sim F_i$, $d_t(i) \sim D_i$
- $\kappa \in [0, 1]$ is transformation factor such that $r_t \in [0, 1]$, $f_t(i) \in [0, \kappa]$
- we set the domain of $d_t(i)$ is $\mathbb{N} \cup \{\infty\}$

$$\mathcal{R}_T = \max_i \mathbb{E}[\Sigma_{t=1}^T (r_t(i) + f_t(i))] - \mathbb{E}[\Sigma_{t=1}^T r_t(a_t) + f_t(a_t)]$$
$$= (1 + \kappa) \times (T\mu_{i^*} - \mathbb{E}[\Sigma_{t=1}^T \mu_{a_t}]) = (1 + \kappa) \times \mathbb{E}[\Sigma_{t=1}^T \Delta_{a_t}]$$

- UCB
- successive elimination
- Exp3
- FTRL
- wrapper

- skipping
- double trick
- drift
- intermidiate observation: Delayed Bandits: When Do Intermediate Observations Help? (2023)