Bandits with Delayed, Aggregated Anonymous Feedback

CS728

Aniruddh Pramod Rahul Jha Rohit Verma Sambuddha Chakrabarti

April 22, 2025

Contents

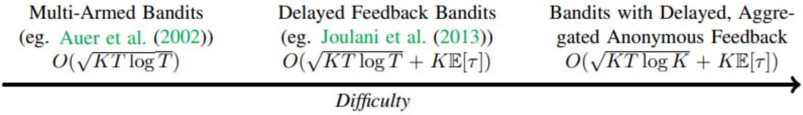
1. Introduction
2. Related Work
3. Problem Definition
4. Techniques and Results
5. The ODAAF Algorithm
6. Regret Bounds
7. Main Result

Introduction: Bandits with Delayed, Aggregated Anonymous Feedback

- **Multi-Armed Bandit (MAB) Problem**:
  - Sequential decision-making framework capturing exploration-exploitation tradeoff
- **Real-world applications with delayed feedback**:
  - Online advertising: conversions happen after ad is shown
  - Personalized treatment: health effects appear after delay
  - Content design: website traffic changes after design updates
- **New challenge**: Delayed, Aggregated Anonymous Feedback (MABDAAF)
  - Rewards are delayed by random time τ
  - Only sum of arriving rewards is observed
  - Cannot determine which arm generated which reward

Related Work

- **Non-anonymous delayed feedback bandits**:
  - **Joulani et al. (2013)**: Showed regret increases only additively by factor of expected delay
  - **Mandel et al. (2015)**: Queue-based methods for bounded delays
  - Key result: $O(\sqrt{KT \log T} + K\mathbb{E}[\tau])$ regret bound
- **Other delayed feedback variants**:
  - **Neu et al. (2010)**, **Cesa-Bianchi et al. (2016)**: Non-stochastic bandits with constant delays
  - **Dudik et al. (2011)**: Stochastic contextual bandits with constant delay
  - **Desautels et al. (2014)**: Gaussian Process bandits with bounded delay
  - **Vernade et al. (2017)**: Delayed Bernoulli bandits with censoring
- **Problem difficulty hierarchy**:



Multi-Armed Bandits (eg. Auer et al. (2002)) $O(\sqrt{KT \log T})$

Delayed Feedback Bandits (eg. Joulani et al. (2013)) $O(\sqrt{KT \log T} + K\mathbb{E}[\tau])$

Bandits with Delayed, Aggregated Anonymous Feedback $O(\sqrt{KT \log K} + K\mathbb{E}[\tau])$

Difficulty

Problem Definition

- **Multi-Armed Bandit Problem**:
    - Player selects one of $K > 1$ arms sequentially.
    - Each arm $j \in \mathcal{A}$ has:
        - Reward distribution ζ_j (mean μ_j, support [0,1]).
        - Delay distribution δ_j (support $\mathbb{N} = \{0, 1, \ldots\}$).
- **Delayed, Aggregated Anonymous Feedback**:
    - Rewards are delayed and aggregated.
    - At the end of round t, player observes:

$$X_t = \sum_{l=1}^{t} \sum_{j=1}^{K} R_{l,j} \times \mathbb{I}\left\{l + \tau_{l,j} = t, J_l = j\right\}.$$

    - Feedback does not specify which arm generated the rewards.
- **Goal**: Maximize cumulative reward or minimize regret.

Regret and Assumptions

- **Regret Definition**: The pseudo-regret over horizon T:

$$\mathfrak{R}_T = T\mu^* - \sum_{t=1}^{T} \mu_{J_t},\ (\mu^* = max_j\ \mu_j).$$

- **Assumptions on Delay Distribution**:
  - Bounded and Known Expected Delay ($E[\tau]$ is known).
  - Delay with Bounded Support (Support$(\tau) = [d], d > 0$).
  - Bounded Variance ($V(\tau) < v, v > 0$).
- **Challenges**:
  - Missing feedback due to delays.
  - Aggregated feedback obscures the origin of rewards.
- Problem complexity increases compared to standard MAB settings.

Techniques and Results

- **Key insight**: Play arms in long consecutive stretches
  - Rarely switching algorithm based on Improved UCB (Auer & Ortner, 2010)
  - Switches arms only $O(\log T)$ times
  - Gradually eliminates suboptimal arms
- **Technical contributions**:
  - Construct appropriate confidence bounds for delayed, aggregated anonymous feedback
  - Use Freedman's inequality to avoid blow-up by factor of K
  - Combine Doob's optimal skipping theorem with Azuma-Hoeffding inequalities
  - Handle dependencies through careful algorithm design with bridge periods
- **Results**: Algorithm ODAAF achieves regret bounds of:
  - $O(\sqrt{KT \log K} + K\mathbb{E}[\tau] \log T)$ using only knowledge of $\mathbb{E}[\tau]$
  - $O(\sqrt{KT \log K} + K\mathbb{E}[\tau])$ with bounded delays $(d \leq \sqrt{T/K})$
  - $O(\sqrt{KT \log K} + K\mathbb{E}[\tau] + K\mathbb{V}(\tau))$ with known variance

ODAAF Algorithm: Core Strategy

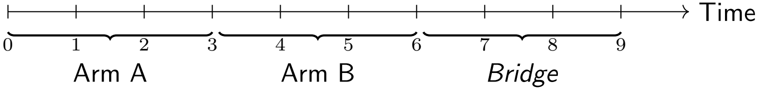
Optimism for Delayed, Aggregated Anonymous Feedback

- A **phase-based arm elimination** algorithm using **rare switching**
- In each phase m:
  1. Play each active arm $j \in A_m$ for $n_m - n_{m-1}$ rounds
  2. Compute empirical mean $\bar{X}_{m,j}$
  3. Eliminate arms if:
  $$\bar{X}_{m,j} + \tilde{\Delta}_m < \max_{j'} \bar{X}_{m,j'}$$
  4. Shrink separation tolerance: $\tilde{\Delta}_{m+1} = \tilde{\Delta}_m / 2$

Bridge Period for Observation Decoupling

- To prevent cross-contamination between phases:
  - A **bridge period** is inserted after each phase
  - A single arm is played for $n_m - n_{m-1}$ steps
  - Feedback from this period is **discarded**

Timeline Visualization



Confidence Bounds and Phase Length

- For reliable elimination, estimate error must satisfy:

$$|\bar{X}_{m,j} - \mu_j| \leq \tilde{\Delta}_m/2$$

- Required number of plays per phase:

$$n_m = \frac{C_1 \log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} + \frac{C_2 m \mathbb{E}[\tau]}{\tilde{\Delta}_m}$$

- Derivation uses:
  - Freedman's inequality (martingale tail)
  - Phase isolation + bounded reward assumption

Regret Bounds

The paper proves the following regret bounds for Bandits with delayed, aggregated feedback. The bounds get more refined as the assumptions on the delay distribution become stronger. We start with the weakest assumption.

Theorem

If only the expected delay $\mathbb{E}(\tau)$ is known, the regret of the algorithm is bounded as

$$\mathbb{E}(R_T) \leq \sum_{j \neq j*} O(\frac{\log(T\Delta_j)^2}{\Delta_j} + \mathbb{E}(\tau)\log(1/\Delta_j))$$

Considering the worst case values of Δ_j we obtain the following

Theorem

$$\mathbb{E}(R_T) \leq O(\sqrt{KT\log K} + K\mathbb{E}(\tau)\log T)$$

Regret Bounds (Contd)

If the delay distribution has a bounded support (i.e is bounded by some $d \geq 0$) then it is relatively easy to bound the number of corrupt observations. The regret bound can then be improved to

Theorem

If the expected delay $\mathbb{E}(\tau)$ is known and the delay distribution is bounded the regret of the algorithm is bounded as

$$\mathbb{E}(R_T) \leq \sum_{j \neq j*} O(\frac{\log(T\Delta_j)^2}{\Delta_j} + \mathbb{E}(\tau) + \min\{d, \frac{\log(T\Delta_j)^2}{\Delta_j}, \mathbb{E}(\tau)\log(1/\Delta_j)\})$$

If $d \leq \mathbb{E}(\tau) + \sqrt{\frac{T \log K}{K}}$ then the bound becomes

Theorem

$$\mathbb{E}(R_T) \leq O(\sqrt{KT \log K} + K\mathbb{E}(\tau))$$

Regret Bounds (Contd)

Finally, if in addition to the expectation, the variance (or an upper bound thereof) $\mathbb{V}(\tau)$ is known, the bounds become

Theorem

$$\mathbb{E}(R_T) \leq \sum_{j \neq j*} O(\frac{\log(T\Delta_j)^2}{\Delta_j} + \mathbb{E}(\tau) + \mathbb{V}(\tau))$$

The corresponding instance independent bound becomes

Theorem

$$\mathbb{E}(R_T) \leq O(\sqrt{KT \log K} + K\mathbb{E}(\tau) + K\mathbb{V}(\tau))$$

It can also be shown that if $\mathbb{E}(\tau) > 1$ the bound becomes

$$\mathbb{E}(R_T) \leq O(\sqrt{KT \log K} + K\mathbb{E}(\tau) + K\mathbb{V}(\tau)/\mathbb{E}(\tau))$$

Preliminaries

1. For each round m, $\tilde{\Delta}_m = \frac{1}{2^m}$, and

$$n_m = \left\lceil \frac{1}{\tilde{\Delta}_m^2} \left(\sqrt{2\log(T\tilde{\Delta}_m^2)} + \sqrt{2\log(T\tilde{\Delta}_m^2) + \frac{8}{3}\tilde{\Delta}_m \log(T\tilde{\Delta}_m^2) + 4\tilde{\Delta}_m(\mathbb{E}[\tau] + 2\mathrm{V}(\tau))} \right)^2 \right\rceil .$$

2. Any arm $j \in \mathcal{A}_j$ is eliminated if $\hat{X}_{m,j} < \max_j \hat{X}_{m,j}$ - $\tilde{\Delta}_m$.
3. At the end of each round, there is a bridge period of length n_m where any active arm is chosen at random and played.
4. Let M_j be the round when arm j is deactivated, and let m_j be the round when it *should* be deactivated: $m_j = \min \left\{ m \,\middle|\, \tilde{\Delta}_m < \frac{\Delta_j}{2} \right\}$
5. This means $\frac{\Delta_j}{4} \leq \tilde{\Delta}_{m_j} \leq \frac{\Delta_j}{2}$

Lemma 24

Statement

Under Assumptions 1 (known $\mathbb{E}[\tau]$) and 3 (known $\mathbb{V}[\tau]$), and with phase lengths n_m from (7), for any arm j and phase m, we have:

$$\mathbb{P}\left(j \in \mathcal{A}_m \text{ and } |\bar{X}_{m,j} - \mu_j| > \tilde{\Delta}_m/2\right) \leq \frac{12}{T\tilde{\Delta}_m^2}$$

Proof Strategy: Decompose observation error into 4 terms:

$$\bar{X}_{m,j} - \mu_j = \frac{1}{n_m}(\text{Term I} + \text{Term II} + \text{Term III} + \text{Term IV})$$

- Term I: Immediate rewards — estimation noise
- Term II: Cross-phase delays — reward inflow
- Term III: Within-phase delays — reward outflow
- Term IV: Expectation mismatch — phase bias

Term I: Immediate reward deviation

- Bounded using Chebyshev's inequality [Lemma 25]
- Applied to:

$$\sum_{t \in T_j(m)} (R_{t,J_t} - \mu_j)$$

- Gives:

$$\mathbb{P}\left(|\text{Term I}| > \frac{n_m \tilde{\Delta}_m}{8}\right) \leq \frac{3}{T \tilde{\Delta}_m^2}$$

Terms II & III: Delayed feedback as Martingale Differences

- (Lemma 26) $Z_t = Q_t - \mathbb{E}[Q_t | \mathcal{G}_{t-1}]$, $Z'_t = \mathbb{E}[P_t | \mathcal{G}_{t-1}] - P_t$ form MDS
- (Lemma 27) Variance bounds:

$$\sum \mathbb{E}[Z_t^2 | \mathcal{G}_{t-1}] \leq 2m \mathbb{V}[\tau], \quad \sum \mathbb{E}[{Z'_t}^2 | \mathcal{G}_{t-1}] \leq m \mathbb{V}[\tau]$$

- Applying Freedman inquality

$$\mathbb{P}\left(|\text{Term II} + \text{Term III}| > \frac{n_m \tilde{\Delta}_m}{4}\right) \leq \frac{6}{T \tilde{\Delta}_m^2}$$

Term IV: Bias from expectation adjustment

- Captures net bias from reward mass entering/leaving the observation window
- (Lemma 28) Provides:

$$|\text{Term IV}| \leq 2\mathbb{E}[\tau] + \frac{4\mathbb{V}[\tau]}{n_m}$$

- Controlled via phase design [Eq. (7)]:

$$n_m \geq \frac{C(\mathbb{E}[\tau] + \mathbb{V}[\tau])}{\tilde{\Delta}_m} \Rightarrow |\text{Term IV}| \leq \frac{n_m \tilde{\Delta}_m}{8}$$

Final Error Bound:

$$\mathbb{P}\left(|\bar{X}_{m,j} - \mu_j| > \tilde{\Delta}_m/2\right) \leq \frac{3+6+3}{T\tilde{\Delta}_m^2} = \frac{12}{T\tilde{\Delta}_m^2}$$

Result: High-probability concentration of empirical mean per phase.

Lemma 29

Statement

For any suboptimal arm j, if $j^* \in \mathcal{A}_{m_j}$, then the probability arm j is not eliminated by round m_j satisfies: $\mathbb{P}\left(M_j > m_j \ \wedge \ M^* \geq m_j\right) \leq \frac{24}{T\tilde{\Delta}^2_{m_j}}$

- Define, $E = \left\{\bar{X}_{m_j,j} \leq \mu_j + w_{m_j}\right\}$ and $H = \left\{\bar{X}_{m_j,j^*} > \mu^* - w_{m_j}\right\}$, ie. the events where the estimate of suboptimal arm is within its 'UCB' and estimate for the optimal arm is within its 'LCB'.
- If E and H occur together, we have, $\bar{X}_{m_j,j} \leq \mu_j^* - \Delta_j + w_{m_j} \leq \bar{X}_{m_j,j^*} - \tilde{\Delta}_{m_j}$, so j gets eliminated.
- So if $M^* > m_j$, we have $M_j \leq m_j$. So if we want both conditions to happen, then either E or H doesn't happen. $\mathbb{P}\Big(M_j > m_j \text{ and } M^* \geq m_j\Big) \leq \mathbb{P}\Big(E^c \cap \{j \in \mathcal{A}_{m_j}\}\Big) + \mathbb{P}\Big(H^c \cap \{j^* \in \mathcal{A}_{m_j}\}\Big)$
- Finally we use Lemma 24, which says that under the given assumptions, for any arm j and phase m, with probability atleast $1 - \frac{12}{T\tilde{\Delta_m^2}}$, either $j \notin \mathcal{A}_m$ or $\bar{X_{m_j}} - \mu_j \leq \tilde{\Delta_m}/2$ to get the result.

Lemma 30

Statement

The probability that optimal arm j^* is eliminated in round $m < \infty$ by suboptimal arm j is bounded by $\mathbb{P}(F_j(m)) \leq \frac{24}{T\tilde{\Delta}_m^2}$

- First note that $F_j(m) = j, j^* \in \mathcal{A}_m$ and $\bar{X}_{m,j} - w_m > \bar{X}_{m,j^*} + w_m$
- Next, if again E and H occur again in round m, $\bar{X}_{m,j} - \tilde{\Delta}_m \leq \bar{X}_{m,j^*}$
- So we again need $E^c \cup H^c$, whose probability is again bounded by Lemma 24

$$\begin{aligned}
\mathbb{P}(F_j(m)) &\leq \mathbb{P}((E^c \cup H^c) \cap (j, j^* \in \mathcal{A}_m)) \\
&\leq \mathbb{P}(E^c \cap (j \in \mathcal{A}_m)) + \mathbb{P}(H^c \cap (j^* \in \mathcal{A}_m)) \\
&\leq 2 \times \frac{12}{T\tilde{\Delta}_m^2} = \frac{24}{T\tilde{\Delta}_m^2}
\end{aligned}$$

Proof

Theorem

If the expected delay and a bound on the variance of the delay $V(\tau)$ is known, the instance dependent upper bound is

$$\mathbb{E}(R_T) \leq \sum_{j \neq j*} O(\frac{\log(T\Delta_j)^2}{\Delta_j} + \mathbb{E}(\tau) + \mathbb{V}(\tau))$$

Proof: Let M^* denote the round where the optimal arm is eliminated, and let $R^i(T)$ be the regret contributed by arm i. The total regret then is $R(T) = \Sigma R^i(T)$. Now we can break the expression of expected regret as

$$\mathbb{E}[R(T)] = \mathbb{E}[\Sigma_i R^i(T)\mathbb{I}\{M^* \geq m_i\}] + \mathbb{E}[\Sigma_i R^i(T)\mathbb{I}\{M^* < m_i\}]$$

We now provide separate bounds for the two terms

Bounding First Term

We begin by rewriting the first term so that we can use the first lemma

$$\mathbb{E}[\Sigma_i R^i(T)\mathbb{I}\{M^* \geq m_i\}] = \mathbb{E}[\Sigma_i R^i(T)\mathbb{I}\{M^* \geq m_i\}\mathbb{I}\{M_i > m_i\}] + \quad (1)$$

$$\mathbb{E}[\Sigma_i R^i(T)\mathbb{I}\{M^* \geq m_i\}\mathbb{I}\{M_i \leq m_i\}] \quad (2)$$

$$\leq \Sigma_i \Delta_i T \mathbb{P}(M^* \geq m_i, M_i > m_i) + \Sigma_i 2\Delta_i n_{m_i,i} \quad (3)$$

$$\leq \Sigma_i(\Delta_i T)\frac{24}{T\tilde{\Delta}_m^2} + \Sigma_i 2\Delta_i n_{m_i,i} \leq \Sigma_i(2\Delta_i n_{m_i,i} + \frac{384}{\Delta_i}) \quad (4)$$

The factor 2 in (3) comes to account for the possibility that we choose a suboptimal arm to play during the bridge period.

Bounding Second Term

First we notice that for the optimal arm j^* to be eliminated in round m by a suboptimal arm with $m_j < m$, arm j both j and j^* must be active in round m_j. We therefore have

$$\mathbb{P}(M^* = m) \leq \Sigma_{i:m_i<m}\mathbb{P}(M_i > m_i, M^* \geq m_j) + \Sigma_{i:m_i \geq m}\mathbb{P}(F_j(m))$$

. We also let $m_{\max} = \max_{j \neq j^*} m_j$

$$\begin{aligned}\mathbb{E}[\Sigma_i R^i(T)\mathbb{I}\{M^* < m_i\}] &= \mathbb{E}[\Sigma_{m=1}^{m_{\max}}\Sigma_{i:m<m_i} R^i(T)\mathbb{I}\{M^* = m\}] \\ &= \Sigma_{m=1}^{m_{\max}}\mathbb{E}[\mathbb{I}\{M^* = m\}\Sigma_{i:m<m_i} N_i\Delta_i] \\ &\leq \Sigma_{m=1}^{m_{\max}}\mathbb{E}[\mathbb{I}\{M^* = m\}T \max_{i:m<m_i} \Delta_i] \\ &\leq \Sigma_{m=1}^{m_{\max}} 4\mathbb{P}(M^* = m)T\tilde{\Delta}_m\end{aligned}$$

Using the value of $P(M^* = m)$ we have $\mathbb{E}[\Sigma_i R^i(T)\mathbb{I}\{M^* < m_i\}] \leq \Sigma_1^K \frac{1920}{\Delta_i}$

Finally

Putting the two bounds together, we have $\mathbb{E}[R(T)] \leq \Sigma_1^K (\frac{1920}{\Delta_i} + 2\Delta_i n_{m_i,i})$. Using the result $(a+b)^2 \leq 2(a^2+b^2)$ for positive a, b we also have the following bound on

$$n_{m_i,i} \leq 1 + \frac{128 L_{m_i,T}}{\Delta_i^2} + \frac{32 L_{m_i,T}}{\Delta_i} + \frac{32\mathbb{E}[\tau]}{\Delta_i} + \frac{64\mathrm{V}(\tau)}{\Delta_i}.$$

Plugging this in, the final bound is

$$\mathbb{E}[R(T)] \leq \sum_{i=1}^{K} \left(\frac{256 \log(T\Delta_i^2)}{\Delta_i} + 64\mathbb{E}[\tau] + 128\mathrm{V}(\tau) + \frac{1920}{\Delta_i} + 64\log(T) + 2\Delta_i \right).$$

This proves the theorem. Now, to get to the instance independent bound

$$\mathbb{E}(R_T) \leq O(\sqrt{KT \log K} + K\mathbb{E}(\tau) + K\mathbb{V}(\tau))$$

consider the following.

Instance Independent Bound

In $\mathbb{E}(R_T) \leq \sum_{j \neq j*} O(\frac{\log(T\Delta_j)^2}{\Delta_j} + \mathbb{E}(\tau) + \mathbb{V}(\tau))$, the only term depending on Δ_j is $\frac{\log(T\Delta_j)^2}{\Delta_j}$ which is maximized for $\lambda = \Delta_j = \sqrt{\frac{K \log Ke^2}{T}}$ so by substituting all Δ_j with λ, we have

$$\mathbb{E}(R_T) \leq CK \log(T\lambda^2)/\lambda + DK(\mathbb{E}(\tau) + \mathbb{V}(\tau))$$

Which becomes $O(\sqrt{KT \log K} + K\mathbb{E}(\tau) + K\mathbb{V}(\tau))$.