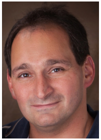
# Adapting to Delays and Data in Adversarial Multi-Armed Bandits



András György

Pooria Joulani

DeepMind

ICML 2021

# Delayed feedback in multi-armed bandit problems

Feedback is often delayed in real-world online learning applications, e.g.,

- recommender systems and web advertisements;
- adaptive clinical trials/optimizing for long-term engagements.

Several works about delayed feedback, e.g., in the adversarial setting,

- Full information (Weinberger and Ordentlich, 2002; Joulani et al., 2013, 2016; Quanrud and Khashabi, 2015)
- Bandit feedback (Neu et al., 2014; Cesa-Bianchi et al., 2016, 2019; Thune et al., 2019; Bistritz et al., 2019; Zimmert and Seldin, 2019, 2020)

This work:

- Fully delay-adaptive version of Exp3 with a remarkably simple proof technique.
- First delay-adaptive method with a high-probability regret bound (based on Exp3-IX).
- First delay- and data-adaptive method.

# Adversarial bandit problem with delayed feedback

Protocol: For $t = 1, 2, \ldots$

- Learner chooses an action $A_t \in [K]$;
- Suffers loss $\ell_{t,A_t}$;
  - loss is revealed after delay $d_t$, in round $t + d_t$;
- Observes feedback $(s, A_s, \ell_{s,A_s})$ for all $s$ with $s + d_s = t$.

Goal: Minimize regret

$$R_T(A^\star) = \sum_{t=1}^{T} \mathbb{E}\left[\ell_{t,A_t}\right] - \sum_{t=1}^{T} \ell_{t,A^\star}$$

where $A^\star = \operatorname{argmin}_{a\in[K]} \sum_{t=1}^{T} \ell_{t,a}$, the optimal action in hindsight.

Assumptions: Loss sequence $\ell_1, \ldots, \ell_T$ and delay sequence $d_1, \ldots, d_T$ are selected in advance.

# Adversarial bandit problem with delayed feedback

Regret in the delayed setting with bandit feedback

- Constant delay: $d_t = d$ for all $t \in [T]$ (Cesa-Bianchi et al., 2016, 2019)

$$R_T = O\left(\sqrt{dT \log K + KT \log K}\right).$$

- Arbitrary delays (Zimmert and Seldin, 2019, 2020)

$$R_T = O\left(\sqrt{D \log K + KT}\right),$$

where $D = \sum_{t=1}^{T} d_t$ is the cumulative delay.

Works for an a priori unknown $D$!

**Question:**

How to adapt Exp3 (in a simple way) to work with an unknown $D$?

- several unsuccessful attempts for adaptation (e.g., Thune et al., 2019; Bistritz et al., 2019).

# The Delay-Adaptive Exp3 Algorithm (DAda-Exp3)

## Delay-adaptive Exp3 (without exploration)

- Loss estimate (importance-weighted): $\hat{\ell}_{t,i} = \dfrac{\ell_{t,i}\mathbb{I}\left[A_t = i\right]}{p_{t,i}}$.
- Action distribution: $p_{t,i} \sim \exp\left(-\eta_t \sum_{s:s+d_s<t} \hat{\ell}_{s,i}\right)$.
  - uses only observed losses

## Analysis

- Non-delayed cheating algorithm: $\tilde{p}_{t,i} \sim \exp\left(-\eta_t \sum_{s=1}^{t} \hat{\ell}_{s,i}\right)$.
- Number of missing feedbacks: $\tau_t = \sum_{s=1}^{t-1} \mathbb{I}\left[s + d_s \geq t\right]$ (note: $\sum_{t=1}^{T} \tau_t = D$).

## Regret bound

$$R_T \leq \underbrace{\eta_T^{-1} \log(K)}_{\text{cheating regret}} + \sum_{t=1}^{T} \underbrace{\eta_t(\tau_t + K)}_{\text{using } p_t \text{ instead of } \tilde{p}_t}$$

# The Delay-Adaptive Exp3 Algorithm (DAda-Exp3)

## Delay-adaptive Exp3 (without exploration)

- Loss estimate (importance-weighted): $\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{I}[A_t = i]}{p_{t,i}}$.
- Action distribution: $p_{t,i} \sim \exp\left(-\eta_t \sum_{s:s+d_s<t} \hat{\ell}_{s,i}\right)$.
  - uses only observed losses

## Analysis

- Non-delayed cheating algorithm: $\tilde{p}_{t,i} \sim \exp\left(-\eta_t \sum_{s=1}^{t} \hat{\ell}_{s,i}\right)$.
- Number of missing feedbacks: $\tau_t = \sum_{s=1}^{t-1} \mathbb{I}[s + d_s \geq t]$ (note: $\sum_{t=1}^{T} \tau_t = D$).

## Regret bound

$$R_T \leq 3\sqrt{\log(K)\,(TK + D)} \text{ for } \eta_t = \sqrt{\frac{\log(K)}{tK + \sum_{s=1}^{t} \tau_s}}.$$

# Variants of DAda-Exp3

High-probability version:

- Implicit exploration (Neu, 2015): $\hat{\ell}_{t,i} = \dfrac{\ell_{t,i}\mathbb{I}\left[A_t = i\right]}{p_{t,i} + \eta_t}$.

Skipping bound:

- Skip round $s$ if $d_s$ proves to be too large (Zimmert and Seldin, 2019, 2020; Thune et al., 2019).
- Regret bound:

$$R_T = O\left(\sqrt{KT\log(K)} + \min_{R\subset[T]}\left\{|R| + \sqrt{D_{\bar{R}}\log(K)}\right\}\right)$$

(guarantees both in expectation and with high-probability).

  - $R$: arbitrary set of rounds.
  - $D_{\bar{R}} = \sum_{t\notin R} d_t$: cumulative delay for rounds not in $R$.

# Delay- and Data-Adaptive Exp3

- Data-dependent learning rate (based on the full-information technique of Joulani et al., 2016)

Computable: $$\eta_t^{-1} \approx (d_t^\star)^2 + \sqrt{\sum_{i=1}^{K} \sum_{s:s+d_s<t} \sum_{r:s\le r+d_r \le s+d_s} \hat{\ell}_{s,i}\hat{\ell}_{r,i}(p_{r,i}+p_{s,i})}$$

needs a priori knowledge of the maximum delay $d_t^\star = \max_{s\le t} d_s$ (similarly to Thune et al., 2019).

- Implicit exploration: $\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{I}[A_t=i]}{p_{t,i}+\eta_t}$.

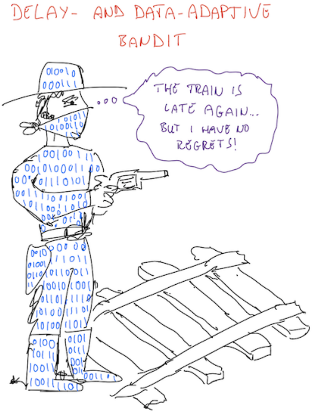
## Regret bound

$$R_T = \tilde{O}\left(d_T^\star + \sqrt{\log(K)\left(d_T^\star L_{T,A^\star} + \sum_{i=1}^{K} L_{T,i}\right)}\right)$$

where $L_{T,i} = \sum_{t=1}^{T} \ell_{t,i}$.

- Similar to the data-dependent bound of Exp3.
- Can be much smaller than $\tilde{O}\left(\sqrt{\log(K)(D+KT)}\right)$.

# For more details and open problems, visit our poster!



# References I


I. Bistritz, Z. Zhou, X. Chen, N. Bambos, and J. Blanchet. Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems 32*, pages 11349–11358. 2019.

N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora. Delay and cooperation in nonstochastic bandits. In *29th Annual Conference on Learning Theory*, pages 605–622, 2016.

N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20(17):1–38, 2019.

P. Joulani, A. György, and C. Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, 2013. (extended arXiv version : http://arxiv.org/abs/1306.0686).

P. Joulani, A. György, and C. Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1744–1750, 2016.

G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3168–3176, 2015.

# References II


G. Neu, A. György, C. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, March 2014.

K. Quanrud and D. Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems 28*, pages 1270–1278. 2015.

T. S. Thune, N. Cesa-Bianchi, and Y. Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems 32*, pages 6541–6550. 2019.

M. J. Weinberger and E. Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, September 2002.

J. Zimmert and Y. Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. *arXiv preprint:1910.06054*, 2019.

J. Zimmert and Y. Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 3285–3294, 2020.