On the Minimax Regret for Online Learning with Feedback Graphs

Khaled Eldowa[1], Emmanuel Esposito[1,4], Tommaso Cesari[2], Nicolò Cesa-Bianchi[1,3]

[1]Università degli Studi di Milano, [2]University of Ottawa

[3]Politecnico di Milano, [4]Istituto Italiano di Tecnologia

Contents

Problem Setting

Problem setting: basic ingredients

- The learner faces an action set $V = [K]$

Problem setting: basic ingredients

- The learner faces an action set $V = [K]$
- A graph $G = (V, E)$ over the actions is provided

Problem setting: basic ingredients

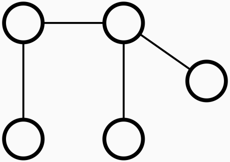
- The learner faces an action set $V = [K]$
- A graph $G = (V, E)$ over the actions is provided
- The player interacts with the environment in a series of T rounds

Problem setting: basic ingredients

- The learner faces an action set $V = [K]$
- A graph $G = (V, E)$ over the actions is provided
- The player interacts with the environment in a series of T rounds
- At the start, the environment (secretly) picks a sequence of losses $(\ell_t)_{t \in [T]}$, where $\ell_t \colon V \to [0, 1]$

Problem setting: interaction protocol and objective

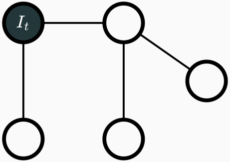
For $t = 1 \ldots T$:



Problem setting: interaction protocol and objective

For $t = 1 \dots T$:

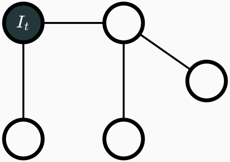
- the learner picks (possibly at random) an action $I_t \in V$



Problem setting: interaction protocol and objective

For $t = 1 \dots T$:

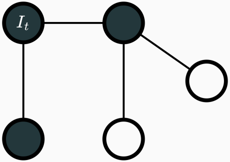
- the learner picks (possibly at random) an action $I_t \in V$
- the learner suffers and observes $\ell_t(I_t)$



Problem setting: interaction protocol and objective

For $t = 1 \dots T$:

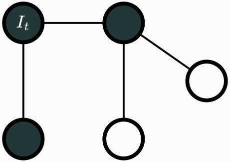
- the learner picks (possibly at random) an action $I_t \in V$
- the learner suffers and observes $\ell_t(I_t)$
- the learner observes the losses of the actions in $N_G(I_t)$ (the neighbourhood of I_t in G)



Problem setting: interaction protocol and objective

For $t = 1 \dots T$:

- the learner picks (possibly at random) an action $I_t \in V$
- the learner suffers and observes $\ell_t(I_t)$
- the learner observes the losses of the actions in $N_G(I_t)$ (the neighbourhood of I_t in G)



The objective is to minimize the regret:

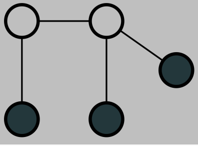
$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i)$$

State of the Art

State of the art

- The minimax regret is the lowest achievable regret of any strategy against its worst case environment
- For a given graph, the best known[1] upper bound is of order $\sqrt{\alpha T \ln K}$

The independence number $\alpha(G)$ is the cardinality of the largest set of nodes no two of which are neighbours



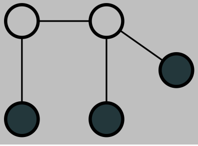
- The best known[1] lower bound is of order $\sqrt{\alpha T}$

[1] Alon et al., 2017.

State of the art

- The minimax regret is the lowest achievable regret of any strategy against its worst case environment
- For a given graph, the best known[1] upper bound is of order $\sqrt{\alpha T \ln K}$

The independence number $\alpha(G)$ is the cardinality of the largest set of nodes no two of which are neighbours



- The best known[1] lower bound is of order $\sqrt{\alpha T}$

[1] Alon et al., 2017.

State of the art: special cases

However, we know that

- for bandits ($\alpha = K$): the minimax regret is[2] $\Theta\big(\sqrt{KT}\big)$
- for experts ($\alpha = 1$): the minimax regret is[3] $\Theta\big(\sqrt{T \ln K}\big)$

[2] Auer et al., 1995; Audibert and Bubeck, 2009.
[3] Cesa-Bianchi et al., 1993.

State of the art: special cases

However, we know that

- for bandits ($\alpha = K$): the minimax regret is[2] $\Theta(\sqrt{KT})$
- for experts ($\alpha = 1$): the minimax regret is[3] $\Theta(\sqrt{T \ln K})$

What about intermediate cases?

[2] Auer et al., 1995; Audibert and Bubeck, 2009.
[3] Cesa-Bianchi et al., 1993.

q-**FTRL**

The FTRL rule

At every round $t = 1, \ldots, T$:

- $\forall i \in [K]$, let $\hat{L}_{t-1}(i) = \sum_{s=1}^{t-1} \hat{\ell}_s(i)$, where $\hat{\ell}_s(i)$ is an estimate of the loss of action i in round s
- Select a distribution over the actions that balances exploitation and exploration/stability:

$$p_t = \operatorname*{arg\,min}_{p \in \Delta_K} \sum_{i=1}^{K} p(i) \hat{L}_{t-1}(i) + \frac{1}{\eta} \psi(p)$$

  where $\psi : \Delta_K \to \mathbb{R}$ is a regularizer
- Draw $I_t \sim p_t$

The FTRL rule

At every round $t = 1, \ldots, T$:

- $\forall i \in [K]$, let $\hat{L}_{t-1}(i) = \sum_{s=1}^{t-1} \hat{\ell}_s(i)$, where $\hat{\ell}_s(i)$ is an estimate of the loss of action i in round s
- Select a distribution over the actions that balances exploitation and exploration/stability:

$$p_t = \operatorname*{arg\,min}_{p \in \Delta_K} \sum_{i=1}^{K} p(i) \hat{L}_{t-1}(i) + \frac{1}{\eta} \psi(p)$$

  where $\psi : \Delta_K \to \mathbb{R}$ is a regularizer
- Draw $I_t \sim p_t$

The importance-weighted estimator

For $i \in [K]$ and $t \in [T]$,

$$\hat{\ell}_t(i) = \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\big\{ I_t \in \{i\} \cup N_G(i) \big\}$$

where $P_t(i) = \mathbb{P}\big(I_t \in \{i\} \cup N_G(i) \mid I_1, \ldots, I_{t-1}\big) = p_t(i) + \sum_{j \in N_G(i)} p_t(j)$

The (negative) q-Tsallis entropy regularizer

For $q \in (0,1)$, define

$$\psi_q(p) = \frac{1}{1-q}\left(1 - \sum_{i=1}^{K} p(i)^q\right)$$

The (negative) q-Tsallis entropy regularizer

For $q \in (0, 1)$, define

$$\psi_q(p) = \frac{1}{1-q}\left(1 - \sum_{i=1}^{K} p(i)^q\right)$$

- with $q = 1/2$, one can achieve $\sqrt{KT}$ regret for bandits
- in the limit as $q \to 1$, we recover the (negative) Shannon entropy, using which we can achieve $\sqrt{\alpha T \ln K}$ regret

The (negative) q-Tsallis entropy regularizer

For $q \in (0, 1)$, define

$$\psi_q(p) = \frac{1}{1-q}\left(1 - \sum_{i=1}^{K} p(i)^q\right)$$

- with $q = 1/2$, one can achieve $\sqrt{KT}$ regret for bandits
- in the limit as $q \to 1$, we recover the (negative) Shannon entropy, using which we can achieve $\sqrt{\alpha T \ln K}$ regret
- what if we choose q as a function of α?

q-FTRL: key lemma

FTRL with regularizer ψ_q (q-FTRL) and the IW estimator satisfies

$$R_T \leq \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q}\,\mathbb{E}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{p_t(i)^{2-q}}{\sum_{j\in\{i\}\cup N_G(i)} p_t(j)}$$

q-FTRL: key lemma

FTRL with regularizer ψ_q (q-FTRL) and the IW estimator satisfies

$$R_T \leq \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q}\,\mathbb{E}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{p_t(i)^{2-q}}{\sum_{j\in\{i\}\cup N_G(i)} p_t(j)}$$

Lemma

Let G be an undirected graph over K nodes. Then, for any $p \in \Delta_{K-1}$ and $q \in [0,1]$

$$\sum_{i=1}^{K}\frac{p_t(i)^{2-q}}{\sum_{j\in\{i\}\cup N_G(i)} p_t(j)} \leq \alpha(G)^q$$

Thus,

$$R_T \leq \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q}\alpha^q T$$

q-FTRL: final bound

Theorem

q-FTRL with

$$q = \frac{1}{2}\left(1 + \frac{\ln(K/\alpha)}{\sqrt{\ln(K/\alpha)^2 + 4} + 2}\right) \in [1/2, 1) \quad \textit{and} \quad \eta = \sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}}$$

satisifes

$$R_T \leq 2\sqrt{e\alpha T\left(2 + \ln(K/\alpha)\right)}$$

Extensions

The uninformed setting

Consider a variant where

[4] Alon et al., 2017.

The uninformed setting

Consider a variant where

- instead of a fixed graph, the environment selects a sequence of graphs $(G_t)_{t \in [T]}$, where $G_t = (V, E_t)$

[4] Alon et al., 2017.

The uninformed setting

Consider a variant where

- instead of a fixed graph, the environment selects a sequence of graphs $(G_t)_{t \in [T]}$, where $G_t = (V, E_t)$
- the learner observes G_t only after selecting I_t

[4] Alon et al., 2017.

The uninformed setting

Consider a variant where

- instead of a fixed graph, the environment selects a sequence of graphs $(G_t)_{t \in [T]}$, where $G_t = (V, E_t)$
- the learner observes G_t only after selecting I_t

State of the art methods[4] can achieve an upper bound of order $\sqrt{\sum_{t=1}^{T} \alpha_t \ln K}$, where $\alpha_t = \alpha(G_t)$

[4] Alon et al., 2017.

The uninformed setting

Consider a variant where

- instead of a fixed graph, the environment selects a sequence of graphs $(G_t)_{t\in[T]}$, where $G_t = (V, E_t)$
- the learner observes G_t only after selecting I_t

State of the art methods[4] can achieve an upper bound of order $\sqrt{\sum_{t=1}^{T} \alpha_t \ln K}$, where $\alpha_t = \alpha(G_t)$

With $\bar{\alpha}_T = \frac{1}{T}\sum_{t=1}^{T} \alpha_t$, utilizing a doubling trick, we can achieve

$$R_T \le c\sqrt{\sum_{t=1}^{T} \alpha_t \left(2 + \ln\left(\frac{K}{\bar{\alpha}_T}\right)\right)} + \log_2 \bar{\alpha}_T$$

[4] Alon et al., 2017.

General strongly observable graphs

- The learner only observes $\ell_t(i)$ for $i \in N_{G_t}(I_t)$
- For every $i \in V$, at least one of the following holds: $i \in N_{G_t}(i)$ or $i \in N_{G_t}(j)$ for all $j \neq i$
- Let $J_t = \left\{ i \in V : i \notin N_{G_t}(i) \text{ and } p_t(i) > 1/2 \right\}$, we can recover the same guarantees using the following loss estimator adapted from (Zimmert and Seldin, 2021)

$$\hat{\ell}_t(i) = \begin{cases} \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\left\{ I_t \in N_{G_t}(i) \right\} & \text{if } i \in V \setminus J_t \\ \frac{\ell_t(i) - 1}{P_t(i)} \mathbb{I}\left\{ I_t \in N_{G_t}(i) \right\} + 1 & \text{if } i \in J_t \end{cases}$$

Lower Bounds

Lower Bounds

Theorem

Pick any α *and* K *such that* $2 \le \alpha \le K$. *Then, for any algorithm and sufficiently large* T*, there exists a sequence of losses and feedback graphs* $G_1, \ldots, G_T$ *such that* $\alpha(G_t) = \alpha$ *for all* $t = 1, \ldots, T$ *and*

$$R_T \ge c\sqrt{\alpha T \log_\alpha K}$$

Lower Bounds

Theorem

Pick any α and K such that $2 \leq \alpha \leq K$. Then, for any algorithm and sufficiently large T, there exists a sequence of losses and feedback graphs $G_1, \ldots, G_T$ such that $\alpha(G_t) = \alpha$ for all $t = 1, \ldots, T$ and

$$R_T \geq c\sqrt{\alpha T \log_\alpha K}$$

Improves upon the $\Omega\big(\sqrt{\alpha T}\big)$ lower bound, however

- not exactly matching
- requires time-varying graphs
- not instance-specific

Lower Bounds

Theorem

Pick any α and K such that $2 \leq \alpha \leq K$. Then, for any algorithm and sufficiently large T, there exists a sequence of losses and feedback graphs $G_1, \ldots, G_T$ such that $\alpha(G_t) = \alpha$ for all $t = 1, \ldots, T$ and

$$R_T \geq c\sqrt{\alpha T \log_\alpha K}$$

Improves upon the $\Omega\big(\sqrt{\alpha T}\big)$ lower bound, however

- not exactly matching
- requires time-varying graphs
- not instance-specific

A more recent work (Chen, He, and Zhang, 2023) shows that for every $\alpha \leq K$ their exists a (fixed) graph G with $\alpha(G) = \alpha$ such that $R_T \geq c\sqrt{\alpha T \ln\big(K/\alpha\big)}$

References i

Cesa-Bianchi, Nicolò et al. (1993). "How to use expert advice". In: *Proceedings of the 25th annual ACM symposium on Theory of Computing*, pp. 382–391.

Auer, Peter et al. (1995). "Gambling in a rigged casino: The adversarial multi-armed bandit problem". In: *Proceedings of IEEE 36th annual foundations of computer science*. IEEE, pp. 322–331.

Audibert, Jean-Yves and Sébastien Bubeck (2009). "Minimax Policies for Adversarial and Stochastic Bandits.". In: *COLT*. Vol. 7, pp. 1–122.

Mannor, Shie and Ohad Shamir (2011). "From bandits to experts: On the value of side-observations". In: *Advances in Neural Information Processing Systems* 24.

References ii

- Alon, Noga et al. (2017). “Nonstochastic Multi-Armed Bandits with Graph-Structured Feedback”. In: *SIAM Journal on Computing* 46.6, pp. 1785–1826.
- Zimmert, Julian and Yevgeny Seldin (2021). “Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits”. In: *The Journal of Machine Learning Research* 22.1, pp. 1310–1358.
- Chen, Houshuang, Yuchen He, and Chihao Zhang (2023). *On Interpolating Experts and Multi-Armed Bandits*. arXiv: 2307.07264 [cs.LG].