Online Learning with Gaussian Payoffs and Side Observations

Yifan Wu[1] **András György**[2] Csaba Szepesvári[1]

[1]Department of Computing Science
University of Alberta

[2]Department of Electrical and Electronic Engineering
Imperial College London

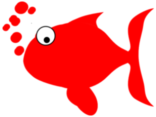
January 12, 2016

Stor-i Multi-armed Bandit Workshop

A Fishy Problem

- Each day, you get to choose a fishing spot.
- Which one to choose?
- Every fish you catch: $+1$ cookies.
- No fish: -10 cookies.
- Fish distribution is i.i.d.
- With some probability, you will see neighboring sites' yield for the day.

The Fishing Game



Choosing a fishing spot: K actions.

$\theta_1, \ldots, \theta_K$: (unknown) mean rewards for the K spots.

For rounds $t = 1, \ldots, T$:

- Choose a fishing spot $I_t \in [K] := \{1, \ldots, K\}$;
- Incur reward $Y_t \in \mathbb{R}$ with mean θ_{I_t};
- Observe $X_t \in \mathbb{R}^K$; noisy reward observations for all the sites $(Y_t = X_{t,I_t})$.

Assumptions

$\mathbb{E}[X_{t,k}] = \theta_k$, and $\mathbb{V}(X_{t,k} | I_t) = \sigma^2_{I_t,k}$ with $\Sigma = (\sigma^2_{i,k})$ known *a priori*.

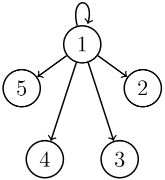
Goal

Minimize expected regret $R_T = T \max_{i \in [K]} \theta_i - \sum_{t=1}^{T} \mathbb{E}[Y_t]$.

Some Interesting Special Cases

- Full information problems: $\sigma_{ij} = \sigma$ for all $i, j \in [K]$.
- Bandits: $\sigma_{ii} = \sigma$ for all $i \in [K]$, $\sigma_{ij} = \infty$ for all $i \neq j$.
- Graph feedback (Alon et al., 2015):
  - Each $i \in [K]$ has $S_i \subset [K]$:

$$\sigma_{i,j} = \begin{cases} \sigma, & \text{if } j \in S_i\,; \\ +\infty, & \text{otherwise}\,. \end{cases}$$



  - Self-observability: $i \in S_i$ for any $i \in [K]$ (Mannor & Shamir, 2011; Caron et al., 2012; Alon et al., 2013; Buccapatnam et al., 2014; Kocák et al., 2014).

Strength: Our single model encompasses all these settings and allows continuous interpolation between them.

How to Compare Algorithms?

Performance Metric

Expected regret $R_T = T \max_{i \in [K]} \theta_i - \sum_{t=1}^{T} \mathbb{E}[Y_t]$.

Minimax Regret:

$$R_T^* = \inf_A \sup_\theta R_T(A, \theta)$$

Typically, $R_T^* = O(T^\alpha)$ with $0 < \alpha < 1$ (polynomial minimax regret), where the constant is a function of (p, r), Θ, but not the individual θ.

Regret Asymptotics:

$\mathcal{A}_s$ = set of algorithms with subpolynomial regret growth, i.e., for any $A \in \mathcal{A}_s$, $\alpha > 0$,

$$R_T(A, \theta) = O(T^\alpha) .$$

Problem-dependent sharp asymptotic regret lower bound: For any $\theta \in \Theta$,

$$\inf_{A \in \mathcal{A}_s} \liminf_{T \to \infty} \frac{R_T(A, \theta)}{\log(T)} = c(\theta) .$$

A Unified Lower Bound

Under our setting with general variance matrix Σ, we have a unified, finite-time, problem-dependent lower bound that recovers all of the existing results.

Lower Bound for Gaussian Case

Idea: Only allow algorithms with bounded worst-case regret over Θ!

Given some $B > 0$, for $i \neq i_1$, let $\Delta_i = \max_j \theta_j - \theta_i$, [1]

$$\epsilon_i = \frac{8\sqrt{e}B}{T} e^{W\left(\frac{\Delta_i T}{16\sqrt{e}B}\right)} + \Delta_i\,, \qquad m_i(\theta, B) = \frac{1}{\epsilon_i^2} \log \frac{T(\epsilon_i - \Delta_i)}{8B}\,.$$

For $i = i_1$, replace Δ_i with Δ_{i_2}. Let

$$C_{\theta,B} = \left\{ c \in C_T^{\mathbb{R}_+} : \sum_{j=1}^{K} \frac{c_j}{\sigma_{ji}^2} \geq m_i(\theta, B) \text{ for all } i \in [K] \right\}.$$

Theorem (Finite-time problem-dependent lower bound)

For any algorithm s.t. $\sup_{\lambda \in \Theta} R_T(\lambda) \leq B$, *any* T *large enough, any* $\theta \in \Theta$,

$$R_T(\theta) \geq b(\theta, B) = \min_{c \in C_{\theta,B}} \sum_{i \neq i_1} c_i \Delta_i\,.$$

[1] $W(\cdot)$ is the Lambert W function satisfying $W(x)e^{W(x)} = x$.

Asymptotic Lower Bound for Graph Feedback

Derived from the work of Graves & Lai (1997):

- Let $\Delta_i = \Delta_i(\theta)$; $\sigma_{i,j} \in \{\sigma, +\infty\}$. Assumption: optimal action is unique; let i_1, i_2 be the index of the best, resp., second best action.

Theorem (Asymptotic lower bound)

For any algorithm $A \in \mathcal{A}_s$, and for any $\theta \in \Theta$,

$$\liminf_{T\to\infty} \frac{R_T(A,\theta)}{\log T} \geq \inf_{c \in C_\theta} \sum_{i \neq i_1} c_i \Delta_i \,,$$

where

$$C_\theta = \left\{ c \in [0,\infty)^K \;:\; \sum_{i : j \in S_i} c_i \geq \frac{2\sigma^2}{\Delta_j^2} \text{ for all } j \neq i_1, \text{ and } \sum_{i : i_1 \in S_i} c_i \geq \frac{2\sigma^2}{\Delta_{i_2}^2} \right\}.$$

Recovering the Asymptotic Lower Bound

Corollary (Finite-time problem-dependent lower bound)

For any algorithm such that $\sup_{\lambda \in \Theta} R_T(\lambda) \leq B$, *we have, for any* $\theta \in \Theta$,

$$R_T(\theta) \geq b(\theta, B) = \min_{c \in C_{\theta,B}} \sum_{i \neq i_1} c_i \Delta_i \,. \qquad (*)$$

- Recall asymptotic lower bound:

$$\liminf_{T \to \infty} \frac{R_T(\theta)}{\log T} \geq \inf_{c \in C_\theta} \sum_{i \neq i_1} c_i \Delta_i \,. \qquad (**)$$

- For any $B = \alpha T^\beta$ with $\alpha > 0$ and $\beta \in (0, 1)$ we have

$$C_{\theta,B} \to \frac{(1-\beta)\log T}{2} C_\theta$$

as $T \to \infty$. Hence, (**) is recovered from (*).

Minimax Lower Bounds (Alon et al., 2015)

Each $i \in [K]$ is associated with an observation set $S_i \subset [K]$: for $j \in S_i$, $\sigma_{ij} = \sigma$; for $j \notin S_i$, $\sigma_{ij} = \infty$.

- Assume Σ is always observable: for all i, there exists j such that $i \in S_j$.
- Σ is strongly observable if all actions are strongly observable.
  - An action i is *strongly observable* if either it is self-observable or is observable under *any* other action. Otherwise, the action is said to be *weakly observable*.
- Σ is weakly observable if it is observable but not strongly observable.

Minimax Lower Bounds for Graph Feedback - Strong Observability

- $\sigma_{i,j} \in \{1, +\infty\}$, $\Theta = [0,1]$; $S_i = \{j : \sigma_{i,j} = \sigma\}$.
- A set $A \subset [K]$ is *independent* in Σ if for any $i \in A$, $S_i \cap A \subset \{i\}$.
  - Choosing $i \in A$ gives no information about any $j \neq i, j \in A$.
- *Independence number* of Σ:

$$\kappa(\Sigma) = \max\{|A| : A \subset [K] \text{ is independent in } \Sigma\}.$$

For $\sup_{\lambda \in \Theta} R_T(\lambda) \leq B$ and $B = \frac{\sigma\sqrt{\kappa(\Sigma)T}}{8\sqrt{e}}$ we have, for any $\theta \in \Theta$, $R_T(\theta) \geq b(\theta, B) \geq B$ for large enough T.

Corollary (Minimax lower bound under strong observability)

For large enough T, for any algorithm, $\sup_{\theta \in \Theta} R_T(\theta) \geq B$.

Recovers bounds of Mannor & Shamir (2011), Alon et al. (2015).

Minimax Lower Bounds for Graph Feedback - Weak Observability

- $\sigma_{i,j} \in \{1, +\infty\}$, $\Theta = [0, 1]$; $S_i = \{j : \sigma_{i,j} = \sigma\}$;
- $A, A' \subset [K]$; A *dominates* A' if for any $j \in A'$ there exists $i \in A$ such that $j \in S_i$;
  - Any $j \in A'$ can be observed through some $i \in A$.
- $\mathcal{W}(\Sigma)$: Set of all weakly observable actions;
- *Weak domination number*:

$$\rho(\Sigma) = \min\{|A| : A \text{ dominates } \mathcal{W}(\Sigma)\}\,.$$

Corollary (Minimax lower bound under weak observability)

Choosing $B = \frac{(\rho(\Sigma)D)^{1/3}(\sigma T)^{2/3}}{73(\log K)^{2/3}}$ *gives* $\sup_{\theta\in\Theta} R_T(\theta) \geq B$ *for any algorithm.*

Recovers bounds of Mannor & Shamir (2011), Alon et al. (2015).

Upcoming Attractions

- Just for feedback graphs;
- Near asymptotically optimal algorithm (new);
- *Single* near-minimax optimal algorithm – with logarithmic asymptotic regret (new).

Asymptotically Optimal Algorithm

Recall

$$C_\theta = \left\{ c \in [0,\infty)^K \; : \; \sum_{i:j\in S_i} c_i \geq \frac{2\sigma^2}{\Delta_j^2} \text{ for all } j \neq i_1, \text{ and } \sum_{i:i_1\in S_i} c_i \geq \frac{2\sigma^2}{\Delta_{i_2}^2} \right\} .$$

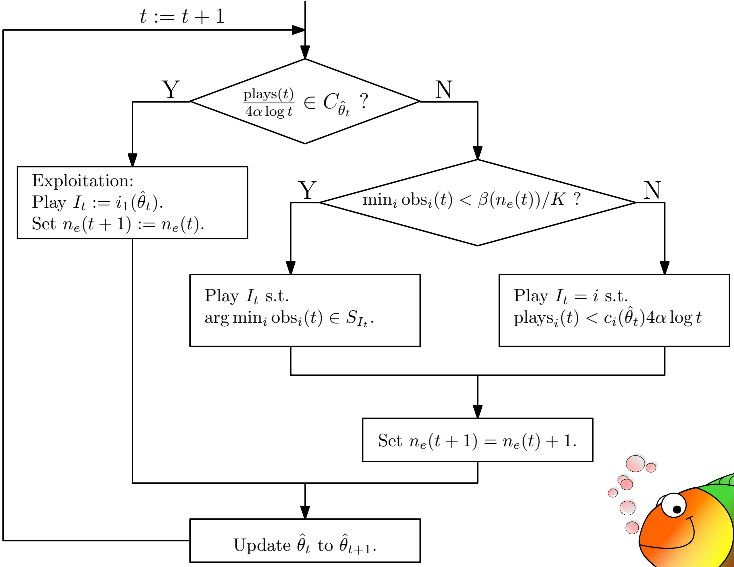
Let $c(\theta) = \operatorname{argmin}_{c\in C_\theta} \sum_{i\neq i_1} c_i \Delta_i$.

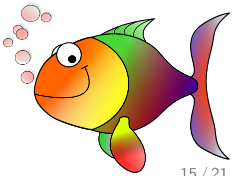
Goal: Find an algorithm that achieves $O(\,(\sum_{i\neq i_1} c_i(\theta)\Delta_i)\log T)$ regret.

(Simple) idea borrowed from Magureanu et al. (2014):

- Use forced exploration to ensure that $c(\theta)$ is well-approximated by $c(\hat{\theta}_t)$ uniformly in time, while paying a constant price in total.
- Targeted minimum number of exploration steps $\beta(\cdot) : \mathbb{N} \to \mathbb{R}$ is chosen to be sublinear.
  - Magureanu et al. (2014)'s linear schedule $\beta(n) = \beta n$ requires that they choose a parameter of their algorithm based on the unknown $\Delta_{\min}$. The sublinear schedule avoids this.

Asymptotically Optimal Algorithm - Pseudocode





Asymptotically Optimal Algorithm - Upper Bound

Upper bound

For any $\alpha > 2$, $\beta(n) = an^b$ with $a \in (0, \frac{1}{2}]$, $b \in (0, 1)$ and for any $\theta \in \Theta$ such that $c(\theta)$ is unique,

$$\limsup_{T \to \infty} \frac{R_T(\theta)}{\log T} \leq 4\alpha \sum_{i \neq i_1} c_i(\theta) \Delta_i \, .$$

Minimax Optimal Algorithm

Successive elimination: maintain a set of possibly optimal actions ("good" actions) until only one action remains.

In each round r,

- Explore all "good actions" by playing only "good actions". (exploitation)
- Due to weak observability, sometimes some actions can only be explored by "bad actions" (exploration-exploitation trade off).
- Use a sublinear function γ to control the exploration using "bad actions".

The idea is similar to the CBP algorithm in Bartók et al. (2014). Here we use a better exploration method to exploit the feedback structure, which leads to the optimal dependence on factors such as $\rho(\Sigma)$ and $\kappa(\Sigma)$.

Minimax Optimal Algorithm - Upper Bound

Theorem

With $\delta = \frac{1}{T}$, for any $\theta \in \Theta$:

- *If Σ is strongly observable,*

$$R_T(\theta) = O\left(\sigma \log K \sqrt{\kappa(\Sigma) T \log T}\right) .$$

- *If Σ is weakly observable,*

$$R_T(\theta) = O\left((\rho(\Sigma) D)^{1/3} (\sigma T)^{2/3} \cdot \sqrt{\log KT}\right) .$$

- *If we view $\Delta_{\min}$ as constant and only consider dependence on T,*

$$R_T(\theta) = O\left(\log^{3/2} T\right) .$$

Conclusions

- Online learning with Gaussian payoffs and side observations;
- Smooth interpolation between full-information and bandit settings;
- First non-asymptotic, problem-dependent lower bounds in regret minimization;
- Algorithms for $\sigma_{i,j} \in \{\sigma, +\infty\}$;
  - Asymptotically near-optimal algorithm;
    - ⋆ First for learning with feedback graphs to do this;
  - Single near minimax algorithm regardless of observability, with poly-logarithmic asymptotic regret;
    - ⋆ First for learning with feedback graphs to do this:
    - ⋆ Mannor & Shamir (2011); Alon et al. (2013) and Alon et al. (2015): No log asymptotic regret, minimax algs.
    - ⋆ Caron et al. (2012) and Buccapatnam et al. (2014): Log asymptotics (with bad dependence on problem parameters), but no near-minimax finite time regret.

Open Problems

- Remove the assumption of a unique optimal arm for the first algorithm;
- Remove the $\log^{1/2} T$ overhead for the second algorithm;
- A single algorithm that achieves both asymptotic and minimax optimal bounds up to constant factors;
  - For bandits, achieved (very) recently (Lattimore, 2015)
- Algorithm for general Σ;
- Algorithm for unknown Σ;
- General tightness of the new lower bound;
- Algorithms for the (general) stochastic partial monitoring setting.

References

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. From bandits to experts: A tale of domination and independence. In *NIPS*, pp. 1610–1618, 2013.

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Online learning with feedback graphs: beyond bandits. In *COLT*, pp. 23–35, 2015.

Bartók, Gábor, Foster, Dean P., Pál, Dávid, Rakhlin, Alexander, and Szepesvári, Csaba. Partial monitoring – classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39:967–997, 2014.

Buccapatnam, Swapna, Eryilmaz, Atilla, and Shroff, Ness B. Stochastic bandits with side observations on networks. *SIGMETRICS Perform. Eval. Rev.*, 42(1):289–300, June 2014.

Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. Leveraging side observations in stochastic bandits. In *UAI*, pp. 142–151, 2012.

Graves, Todd L. and Lai, Tze Leung. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35 (3):715–743, 1997.

Kocák, Tomáš, Neu, Gergely, Valko, Michal, and Munos, Rémi. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 613–621, 2014.

Lattimore, T. Optimally confident UCB: Improved regret for finite-armed bandits. Arxiv preprint, 2015.

Magureanu, S., Combes, R., and Proutiere, A. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *COLT*, pp. 975–999, 2014.

Mannor, S. and Shamir, O. From bandits to experts: on the value of side-observations. In *NIPS*, pp. 684–692, 2011.