Online Learning and Blackwell Approachability with Partial Monitoring : Optimal Convergence Rates

Joon Kwon — Vianney Perchet

CMAP, École polytechnique, Université Paris–Saclay — CMLA, École normale supérieure de Paris–Saclay & Criteo Research

Introduction

- Blackwell approchability is a very general online learning framework, which contains as special cases regret minimization and many of its variants (internal/swap regret, online combinatorial optimization, etc.)
- We consider *partial monitoring*, meaning that the decision maker only observes random signals whose laws may depend on his action and the state of Nature.
- We establish that the worst-case convergence rates are of order $O(T^{-1/2})$ for *outcome-dependent signals* (i.e if the laws of the signals are independent of the actions of the decision maker), and of order $O(T^{-1/3})$ for general signals.

Approachability with full information

- The framework is a repeated decision process between a *decision maker* and *Nature*, with vector-valued payoffs.
- The decision maker (resp. Nature) has finite action set $\mathcal{I}$ (resp. $\mathcal{J}$)
- Denote $\Delta(\mathcal{I})$ (resp. $\Delta(\mathcal{J})$) the set of probability distributions on $\mathcal{I}$ (resp. $\mathcal{J}$)
- Let $\mathbf{g} : \mathcal{I} \times \mathcal{J} \to \mathbb{R}^d$ be a *payoff function*.
- At each stage $t \geqslant 1$,
  1. the decision maker chooses $x_t \in \Delta(\mathcal{I})$;
  2. Nature chooses $y_t \in \Delta(\mathcal{J})$;
  3. actions $i_t \sim x_t$ and $j_t \sim y_t$ are drawn;
  4. the decision maker gets and observes vectorial payoff $g_t := \mathbf{g}(i_t, j_t)$.
- Let $\mathcal{C} \subset \mathbb{R}^d$ be a closed and convex *target set*.

Questions

- Is there an algorithm for the decision maker such that

$$\mathbb{E}\left[\mathbf{d}_2\left(\frac{1}{T}\sum_{t=1}^{T} g_t,\ \mathcal{C}\right)\right] \xrightarrow[T\to+\infty]{} 0$$

  (where $\mathbf{d}_2(\,\cdot\,,\ \mathcal{C})$ denotes the Euclidean distance to $\mathcal{C}$) against any choices $y_1, \dots, y_T \in \Delta(\mathcal{J})$ from Nature ? It depends on $\mathbf{g}$ and $\mathcal{C}$. When this is the case, target set $\mathcal{C}$ is said to be *approachable* by the decision maker.
- Is there a guaranteed (i.e. worst-case) *speed of convergence* for approachable target sets? **Yes**, it is known to be of order $O(T^{-1/2})$ [Bla54, Bla56].

Approachability with partial monitoring

- Let $\mathcal{S}$ be a finite set of *signals*
- Let $\mathbf{s} : \mathcal{I} \times \mathcal{J} \to \Delta(\mathcal{S})$ be the *signal distribution function*
- At each stage $t \geqslant 1$,
  1. the decision maker chooses $x_t \in \Delta(\mathcal{I})$;
  2. Nature chooses $y_t \in \Delta(\mathcal{J})$;
  3. actions $i_t \sim x_t$ and $j_t \sim y_t$ are drawn;
  4. the decision maker gets (**but does not observe**) vectorial payoff $g_t := \mathbf{g}(i_t, j_t)$.
  5. the decision maker observes signal $s_t \sim \mathbf{s}(i_t, j_t)$.

Motivation and examples

Partial monitoring occurs naturally in situations where getting information (i.e. feedback) is costly, leading the decision maker to sometimes choose actions which give limited or partial information.

- In the *apple tasting problem*, for each apple, the decision maker must decide whether it is *good for sale* or *rotten*. The only way to know the actual answer is to open the apple, which is costly, because an opened apple cannot be sold.
- In *automatic subtitling*, the only way for the algorithm to check the quality of his decisions is to ask a human, which is costly.

We give an example with *random* signals.

- To send data in a *congested network*, a computer tries to choose the least congested path. The only feedback is the number of lost packets, which is a random variable whose law depends on the congestion level of the path.

Question and known results

Question

With partial monitoring, what is the worst-case convergence speed (when the target set is approachable)?

The following convergence speeds were previously achieved.
In the special case of regret minimization:

- Early works by [Rus99, MS03]
- [LMS08]: $O(T^{-1/5}\sqrt{\log T})$
- [Per11b]: $O(T^{-1/3})$ (optimal, i.e. worst-case)

In the case of approachability with polytope target sets:

- [Per11a]: $O(T^{-1/(|\mathcal{I}|+3)})$
- [MPS14] and [MPS13]: $O(T^{-1/5})$

Main results

We assume the target set $\mathcal{C} \subset \mathbb{R}^d$ to be a polytope.

Theorem

Let $T \geqslant 1$.

(i) There exists an algorithm for the decision maker such that

$$\mathbb{E}\left[\mathbf{d}_2\left(\frac{1}{T}\sum_{t=1}^{T} g_t,\ \mathcal{C}\right)\right] = O\left(T^{-1/3}\right)$$

for all choices $y_1, \dots, y_T \in \Delta(\mathcal{J})$ from Nature.

(ii) In the case of outcome-dependent signals, i.e for all $j \in \mathcal{J}$

$$\mathbf{s}(\,\cdot\,, j) \text{ is constant,}$$

there exists an algorithm for the decision maker such that

$$\mathbb{E}\left[\mathbf{d}_2\left(\frac{1}{T}\sum_{t=1}^{T} g_t,\ \mathcal{C}\right)\right] = O\left(T^{-1/2}\right)$$

for all choices $y_1, \dots, y_T \in \Delta(\mathcal{J})$ from Nature.

References

[Bla54] David Blackwell.
Controlled random walks.
In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 336–338, 1954.

[Bla56] David Blackwell.
An analog of the minimax theorem for vector payoffs.
Pacific Journal of Mathematics, 6(1):1–8, 1956.

[LMS08] Gábor Lugosi, Shie Mannor, and Gilles Stoltz.
Strategies for prediction under imperfect monitoring.
Mathematics of Operations Research, 33(3):513–528, 2008.

[MPS13] Shie Mannor, Vianney Perchet, and Gilles Stoltz.
A primal condition for approachability with partial monitoring.
Journal of Dynamics and Games, 1(3):447–469, 2013.

[MPS14] Shie Mannor, Vianney Perchet, and Gilles Stoltz.
Set-valued approachability and online learning with partial monitoring.
The Journal of Machine Learning Research, 15(1):3247–3295, 2014.

[MS03] Shie Mannor and Nahum Shimkin.
On-line learning with imperfect monitoring.
In *Learning Theory and Kernel Machines*, pages 552–566. Springer, 2003.

[Per11a] Vianney Perchet.
Approachability of convex sets in games with partial monitoring.
Journal of Optimization Theory and Applications, 149(3):665–677, 2011.

[Per11b] Vianney Perchet.
Internal regret with partial monitoring: Calibration-based optimal algorithms.
The Journal of Machine Learning Research, 12:1893–1921, 2011.

[Rus99] Aldo Rustichini.
Minimizing regret: The general case.
Games and Economic Behavior, 29(1):224–243, 1999.