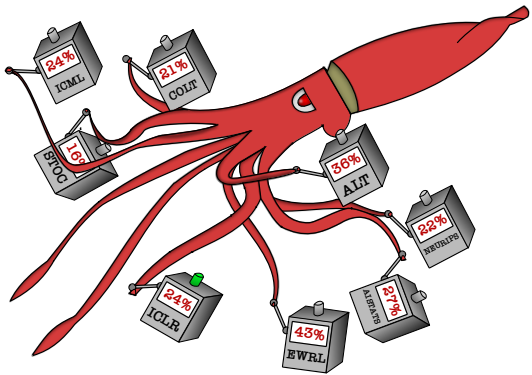
Partial Monitoring – Infinite Outcomes and Rustichini's Regret

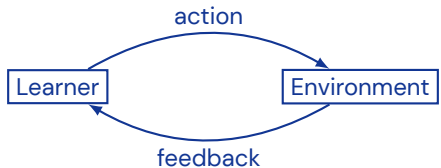
Tor Lattimore

DeepMind, London



Partial Monitoring

- A general framework for sequential decision making
- POMDPs without planning; or
- Bandits with complicated feedback structures
- Loss not **directly** observed



Partial monitoring mathematically

- A game is defined by a pair of functions
- A loss function: $\mathcal{L} : \mathcal{A} \times \mathcal{Z} \to [0, 1]$
- A signal function: $\mathcal{S} : \mathcal{A} \times \mathcal{Z} \to \Sigma$
- $\mathcal{A}$ is the action set, $\mathcal{Z}$ is the latent space and Σ is the space of possible signals
- **Important** $\mathcal{S}$ and $\mathcal{L}$ are known

Interaction protocol

- Game is played over n rounds
- Adversary **secretly** chooses distributions $(x_t)_{t=1}^n$ from $\Delta(\mathcal{Z})$
- For rounds $t = 1$ to n, learner chooses action $a_t \in \mathcal{A}$ based on history
- Observes feedback/signal $\sigma_t = \mathcal{S}(a_t, z_t)$ with $z_t \sim x_t$
- Suffers loss $\mathcal{L}(a_t, z_t)$ but this is **not directly observed**

Extensions

- $\mathcal{L}(\pi, x) = \mathbb{E}_{(a,z)\sim\pi\otimes x}[\mathcal{L}(a, z)]$
- $\mathcal{S}(a, x)$ is the law of $\mathcal{S}(a, z)$ when $z \sim x \in \Delta(\mathcal{Z})$

Examples

- Prediction with expert advice
- Bandits
- Bandits with graph feedback
- Linear bandits
- Convex bandits
- ...
- Apple tasing
- Dynamic pricing
- Spam filtering
- Matrix games

Notions of regret

Standard definition of regret compares learner's cumulative loss to the best single action in hindsight:

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \mathcal{L}(\pi_t, x_t)\right] - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \mathcal{L}(a, x_t)$$

Expectation comes from randomisation of the learner (if any)

Minimax regret is

$$R_n^\star = \inf_{\text{policies}} \sup_{\text{adversary}} R_n$$

Question How does the minimax regret depend on the loss and signal functions? And what algorithms make the regret small

This isn't the only notion of regret. Sometimes we compare to different baseline (usually stronger). In partial monitoring we sometimes need a **weaker** baseline.

Hopeless games

Matching pennies in the dark (Perchet, 2011)

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \mathcal{S} = \begin{pmatrix} \bot & \bot \\ \bot & \bot \end{pmatrix}$$

Learner never observes useful feedback and regret is linear: $R_n^\star = \Omega(n)$

This does not seem like a very fair game

Rustichini's regret

Rustichini (1999) compared the learner to the best policy (distribution over actions) given knowledge of the **average observable signal distribution**

Given $x \in \Delta(\mathcal{Z})$ and $a \in \mathcal{A}$, let $\mathcal{S}(a, x)$ be the law of $\mathcal{S}(a, z)$ when $z \sim x$

Define an equivalence relation $\mathbf{R}$ on $\Delta(\mathcal{Z})$ by $x\mathbf{R}y$ if $\mathcal{S}(a, x) = \mathcal{S}(a, y)$ for all $a \in \mathcal{A}$

Define a function $\mathcal{V} : \Delta(\mathcal{A}) \times \Delta(\mathcal{Z}) \to \mathbb{R}$ by

$$\mathcal{V}(\pi, x) = \max_{y\mathbf{R}x} \mathcal{L}(\pi, y)$$

Rustichini's regret is

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \mathcal{V}(\pi_t, x_t)\right] - n\mathcal{V}_\star\left(\frac{1}{n}\sum_{t=1}^{n} x_t\right) \qquad \mathcal{V}_\star(x) = \min_{\pi} \mathcal{V}(\pi, x)$$

Note $x \mapsto \mathcal{V}(\pi, x)$ and $x \mapsto \mathcal{V}_\star(x)$ are concave and $\mathcal{V}_\star$ is piecewise linear

Conventions

Standard setting

$\mathcal{V}(\pi, x) = \mathcal{L}(\pi, x)$ and $\mathcal{Z}$ is arbitrary

Rustichini setting

$\mathcal{V}(\pi, x) = \mathbf{max}_{y \mathbf{R} x}\, \mathcal{L}(\pi, y)$ and $\mathcal{Z}$ is **finite**

Regret/minimax regret are

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \mathcal{V}(\pi_t, x_t)\right] - n\mathcal{V}_\star\left(\frac{1}{n}\sum_{t=1}^{n} x_t\right) \qquad R_n^\star = \min_{\text{policies}} \sup_{\text{adversary}} R_n$$

Important special case if $x_t = x$ for all t

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \Delta(\pi_t, x)\right] \qquad \Delta(\pi, x) = \mathcal{V}(\pi, x) - \mathcal{V}_\star(x)$$

What do we know ...

... about the dependence on the horizon of the minimax regret?

- Many examples fully understood: PwEA, bandits, linear bandits, convex bandits and many many more
- When $\mathcal{Z}$ is finite and the regret is the standard definition, then

$$R_n^\star \in \{0, \Theta(n^{1/2}), \Theta(n^{2/3}), \Theta(n)\}$$

  This is the classification theorem (Bartók et al., 2014)
- When $\mathcal{Z}$ is finite and the regret is the Rustichini definition, then $R_n^\star = O(n^{2/3})$ (Kwon and Perchet, 2017)

What's new?

- A characterisation of the regret in the standard setting for infinite $\mathcal{Z}$
- A characterisation of the regret in the Rustichini setting for finite $\mathcal{Z}$

Caveat: Finite action sets only, horizon dependence only

Information ratio

- Suppose that $x \in \Delta(\mathcal{Z})$ is sampled from known distribution μ ('prior')
- Learner chooses a distribution $\pi \in \Delta(\mathcal{A})$
- $a \sim \pi$ and $z \sim x$ and observation is $\sigma = \mathcal{S}(a, z)$
- The information ratio measures the tradeoff between the information gained by the learner and the regret suffered:

$$\Psi^{\lambda}(\pi, \mu) = \frac{\Delta(\pi, \mu)^{\lambda}}{I(\pi, \mu)} \qquad \Delta(\pi, \mu) = \int_{\Delta(\mathcal{Z})} \Delta(\pi, x) \, \mathbf{d}\mu(x) \qquad \Delta(\pi, x) = \mathcal{V}(\pi, x) - \mathcal{V}_{\star}(x)$$

- $I(\pi, \mu)$ needs to measure how much information is gained by the policy

$$I(\pi, \mu) = I(x; a, \sigma) = \mathbb{E}\left[\sum_{x \in \mathbf{spt}(\mu)} \mu(x) \, \mathbf{KL}(\mathcal{S}(a, x), \mathcal{S}(a, \mu))\right]$$

- Alternative

$$I(\pi, \mu) = \mathbb{E}\left[\sum_{x \in \mathbf{spt}(\mu)} \mathbf{KL}(\mathcal{S}(a, \mu), \mathcal{S}(a, x))\right]$$

Main theorem

The information ratio characterises the minimax regret up to **subpolynomial** factors

Theorem
The minimax λ-information ratio is

$$\Psi_\star^\lambda = \sup_\mu \inf_\pi \Psi^\lambda(\pi, \mu) = \sup_\mu \inf_\pi \frac{\Delta(\pi, \mu)^\lambda}{I(\pi, \mu)}$$

$\lambda \mapsto \Psi_\star^\lambda$ is decreasing because $\Delta(\pi, \mu) \in [0, 1]$

$$\lambda_\star = \mathbf{sup}\{\lambda > 1 : \Psi_\star^\lambda = \infty\}$$

In both the *standard setting* and *Rustichini setting*, the minimax regret satisfies

$$\limsup_{n \to \infty} \frac{\log(R_n^\star)}{\log(n)} = 1 - \frac{1}{\lambda_\star}$$

Information-directed Sampling

- Bayesian algorithm
- Bayesian setting
- Frequentist results via minimax duality

Learner is given a prior ξ on $\Delta(\mathcal{Z})^n$ and $(x_t)_{t=1}^n$ is sampled from ξ

Bayesian regret is

$$BR_n = \mathbb{E}_{(x_t)\sim\xi}[R_n] = \mathbb{E}\left[\sum_{t=1}^n \mathcal{V}(\pi_t, x_t) - n\mathcal{V}_\star\left(\frac{1}{n}\sum_{t=1}^n x_t\right)\right]$$

Information-directed sampling plays π_t to minimise

$$\pi \mapsto \Psi^\lambda(\pi, \mu_t)$$

where μ_t has law $\mathbb{E}_{t-1}[x_t|G]$ and $G = \nabla\mathcal{V}_\star\left(\frac{1}{n}\sum_{t=1}^n x_t\right) \in \mathcal{G} = \mathbf{Im}(\nabla\mathcal{V}_\star)$

$$
\begin{aligned}
BR_n &= \mathbb{E}\left[\sum_{t=1}^{n} \mathcal{V}(\pi_t, x_t) - n\mathcal{V}_\star\left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{E}_{t-1}[\mathcal{V}(\pi_t, x_t) - \mathcal{V}_G(x_t)|G]\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{E}_{t-1}[\mathcal{V}(\pi_t, x_t)|G] - \mathcal{V}_G(\mathbb{E}_{t-1}[x_t|G])\right] && (x \mapsto \mathcal{V}_G(x) \text{ linear}) \\
&\leq \sum_{t=1}^{n} \mathbb{E}\left[\mathcal{V}(\pi_t, \mathbb{E}_{t-1}[x_t|G]) - \mathcal{V}_G(\mathbb{E}_{t-1}[x_t|G])\right] && (\text{concavity of } x \mapsto \mathcal{V}(\pi, x)) \\
&\leq \sum_{t=1}^{n} \mathbb{E}\left[\Delta(\pi_t, \mu_t)\right] && (\text{definition of } \Delta,\ \mathcal{V}_\star \leq \mathcal{V}_G) \\
&= \sum_{t=1}^{n} \mathbb{E}\left[\Psi^\lambda(\pi_t, \mu_t)^{1/\lambda} I(\pi_t, \mu_t)^{1/\lambda}\right] && (\text{definition of information ratio}) \\
&\leq (\Psi_\star^\lambda)^{1/\lambda} \mathbb{E}\left[\sum_{t=1}^{n} I(\pi_t, \mu_t)^{1/\lambda}\right] && (\text{definition of } \pi_t \text{ and } \Psi_\star^\lambda) \\
&\leq (\Psi_\star^\lambda)^{1/\lambda} n^{1-1/\lambda} \mathbb{E}\left[\sum_{t=1}^{n} I(\pi_t, \mu_t)\right]^{1/\lambda} && (\text{Hölder's inequality})
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n} I(\pi_t, \mu_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{g\in\mathcal{G}} \mathbf{KL}(\mathcal{S}(a_t, \mathbb{E}_{t-1}[x]), \mathcal{S}(a_t, \mathbb{E}_{t-1}[x|G=g])\right] && \text{(def. of inf. gain)} \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{g\in\mathcal{G}} \mathbf{log}\left(\frac{\mathbb{P}_{t-1}(\mathcal{S}(a_t, z) = \sigma_t | a_t)}{\mathbb{P}_{t-1}(\mathcal{S}(a_t, z) = \sigma_t | G=g, a_t)}\right)\right] && \text{(def. of KL)} \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{g\in\mathcal{G}} \mathbf{log}\left(\frac{\mathbb{P}_{t-1}(G=g)}{\mathbb{P}_{t-1}(G=g|\mathcal{S}(a_t, z) = \sigma_t, a_t)}\right)\right] && \text{(Bayes' law)} \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{g\in\mathcal{G}} \mathbf{log}\left(\frac{\mathbb{P}_{t-1}(G=g)}{\mathbb{P}_{t}(G=g)}\right)\right] && \text{(def. of } \mathbb{P}_t\text{)} \\
&= \sum_{g\in\mathcal{G}} \mathbb{E}\left[\mathbf{log}\left(\frac{\mathbb{P}_0(G=g)}{P_n(G=g)}\right)\right] && \text{(telescope)} \\
&\leq |\mathcal{G}|\,\mathbf{log}(n)\,. && \text{(ugly tricks)}
\end{aligned}
$$

We're secretly telescoping expected Bregman divergences between $(\mathbb{P}_t(G=\cdot))_{t=1}^n$ with respect to the logarithmic barrier on $|\mathcal{G}|$-simplex

Putting together the last two slides

$$BR_n \leq (\Psi_\star^\lambda)^{1/\lambda} n^{1-1/\lambda} (|\mathcal{G}| \log(n))^{1/\lambda}$$

This holds for any prior, so by minimax theory (L and Szepesvári, 2019)

$$\min_{\text{policy } (x_t)} \max R_n = \min_{\text{policy}} \max_{\text{prior}} BR_n = \max_{\text{prior}} \min_{\text{policy}} BR_n \leq (\Psi_\star^\lambda)^{1/\lambda} n^{1-1/\lambda} (|\mathcal{G}| \log(n))^{1/\lambda}$$

For $\lambda > \lambda_\star = \sup\{\lambda > 1 : \Psi_\star^\lambda = \infty\}$ we have $\Psi_\star^\lambda < \infty$ and hence

$$\limsup_{n \to \infty} \frac{\log(R_n)}{\log(n)} \leq 1 - \frac{1}{\lambda}$$

Take limit as λ tends to $\lambda_\star$ from above gives

$$\limsup_{n \to \infty} \frac{\log(R_n)}{\log(n)} \leq 1 - \frac{1}{\lambda_\star}$$

Not an algorithm. Tools by L and Gyorgy (2021) show that some version of mirror descent with log barrier regularisation achieves the same bound

Lower bounds

Remember $\lambda_\star = \sup\{\lambda > 1 : \Psi_\lambda^\star = \infty\}$ and

$$\Psi^\lambda(\pi,\mu) = \frac{\Delta(\pi,\mu)^\lambda}{I(\pi,\mu)} \quad \Psi_\star^\lambda(\mu) = \min_{\pi\in\Delta(\mathcal{A})} \Psi^\lambda(\pi,\mu) \quad \Psi_\star^\lambda = \sup_\mu \Psi_\star^\lambda(\mu) \quad \Delta_\star(\mu) = \min_{\pi\in\Delta(\mathcal{A})} \Delta(\pi,\mu)$$

WTS: $\limsup_{n\to\infty} \frac{\log(R_n^\star)}{\log(n)} \geq 1 - \frac{1}{\lambda_\star} \iff \limsup_{n\to\infty} \frac{\log(R_n^\star)}{\log(n)} \geq 1 - \frac{1}{\lambda}$ for all $\lambda < \lambda_\star$

1. $\lambda < \lambda_\star$ implies $\Psi_\star^\lambda = \infty$ and hence μ can be chosen so that $\Psi_\star^\lambda(\mu)$ is arbitrarily large
2. Next slide we'll show that $R_n^\star = \Omega\left(\min\left(n\Delta_\star(\mu), \frac{\Psi_\star^\lambda(\mu)}{\Delta_\star(\mu)^{\lambda-1}}\right)\right)$
3. Choosing $n = \Psi_\star^\lambda(\mu)/\Delta_\star(\mu)^\lambda$ $\left[\to \infty \text{ as } \Psi_\star^\lambda(\mu) \to \infty\right]$ yields
$$R_n^\star = \Omega\left(n^{1-1/\lambda}\Psi_\star(\mu)^{1/\lambda}\right)$$
4. Taking logs and limits completes the proof

- Let $\mu \in \Delta(\mathcal{Z})$
- $\mathbb{P}$ is the measure on interaction sequences when the adversary samples $z_t \sim \int x \, \mathbf{d}\mu(x)$
- $\mathbb{P}_x$ is the measure on interaction sequences when the adversary samples $z_t \sim x$

Lemma 1 $\Delta(\pi, \mu) \geq 2^{-\lambda} \Psi_\star^\lambda(\mu) I(\pi, \mu) / \Delta_\star(\mu)^{\lambda-1}$

Lemma 2 If $(\pi_x)_{x \in \mathbf{spt}(\mu)} \in B_\epsilon(\pi)$ for some π, then $\mathbf{max}_{x \in \mathbf{spt}(\mu)} \Delta(\pi_x, x) = \Omega(\Delta_\star(\mu))$.

$$\mathbf{KL}(\mathbb{P}^\tau, \mathbb{P}_x^\tau) = \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbf{KL}(\mathcal{S}(a_t, \mu), \mathcal{S}(a_t, x))\right] \leq \mathbb{E}\left[\sum_{t=1}^{\tau} \sum_{y \in \mathbf{spt}(\mu)} \mathbf{KL}(\mathcal{S}(a_t, \mu), \mathcal{S}(a_t, y))\right] = \mathbb{E}\left[\sum_{t=1}^{\tau} I(\pi_t, \mu)\right]$$

Case 1: $\mathbb{E}[\sum_{t=1}^{n} I(\pi_t, \mu)] = O(\epsilon)$. Then $\mathbf{KL}(\mathbb{P}^n, \mathbb{P}_x^n) = O(\epsilon)$ for all $x \in \mathbf{spt}(\mu)$. Then learner behaves similarly for all $x \in \mathbf{spt}(\mu)$. Regret is $\Omega(n\Delta_\star(\mu))$ by Lemma 2.

Case 2: $\mathbb{E}[\sum_{t=1}^{\tau} I(\pi_t, \mu)] = \Theta(\epsilon)$. By Lemma 2, $\mathbb{E}[\sum_{t=1}^{\tau} \Delta(\pi_t, \mu)] = \Omega(\Psi_\star^\lambda(\mu)/\Delta_\star(\mu)^{\lambda-1})$. Hence there exists some $x \in \mathbf{spt}(\mu)$ such that $\mathbb{E}_x[\sum_{t=1}^{\tau} \Delta(\pi_t, x)] = \Omega(\Psi_\star^\lambda(\mu)/\Delta_\star(\mu)^{\lambda-1})$.

Hence: $R_n = \Omega(\mathbf{min}(n\Delta_\star(\mu), \Psi_\star^\lambda(\mu)/\Delta_\star(\mu)^{\lambda-1}))$

Summary

- The minimax information ratio determines the minimax regret up to subpolynomial factors for all finite action games
- Result holds for both the standard and Rustichini settings
- More or less practical algorithm for small games
- Stochastic games are not easier than adversarial games (in terms of horizon)
- Example finite Rustichini games with weird minimax regret: $R_n^\star = \Theta(n^{4/7})$
- Example infinite standard games with weird minimax regret: $R_n^\star = \Theta(n^p)$ for any $p \in [1/2, 1]$
- Paper: `https://arxiv.org/abs/2202.10997`

Open questions

- Does the result hold for the standard information ratio (no?)
- What to do when you have many actions? (bounds scale with $|\mathcal{A}|$)
- Classification of all finite Rustichini games

$$\lim_{n\to\infty} \frac{\log(R_n^\star)}{\log(n)} \in \{0, 1/2\} \cup \left\{ \frac{2^i}{2^{i+1}-1} : i \in \mathbb{N} \right\}$$

- Any handle on constants?
- High probability bounds
- Any kind of 'general' infinite-outcome games where we can expect efficient algorithms
- What are the best applications of partial monitoring (and Rustichini's regret)

Bibliography

G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

J. Kwon and V. Perchet. Online learning and blackwell approachability with partial monitoring: optimal convergence rates. In *Artificial Intelligence and Statistics*, pages 604–613. PMLR, 2017.

T. L and A. Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

T. L and C. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the 32nd Conference on Learning Theory*, pages 2111–2139, Phoenix, USA, 2019. PMLR.

V. Perchet. Internal regret with partial monitoring: Calibration-based optimal algorithms. *Journal of Machine Learning Research*, 12(6), 2011.

A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1): 224–243, 1999.