Stat 260/CS 294-102. Learning in Sequential Decision Problems.

Peter Bartlett

1. Discrete decision problems with partial monitoring
   - Definition: loss and feedback. Stochastic and adversarial.
   - Examples.
   - Minimax regret: algorithms and lower bounds.

Discrete decision problems with partial monitoring

Example: **Dynamic pricing**

A vendor has products to sell one by one to a stream of customers. To each, she offers the product at a certain price $p_t \in [0.00, 0.01, \ldots, 1.00]$. The customer has in mind a maximal price m_t that he's willing to pay. If $p_t \leq m_t$, the customer buys the product, otherwise he does not (and in neither case does he reveal his maximal price). The loss of the vendor is missed earnings plus a fixed cost per customer:

$$L_{p_t,m_t} = (m_t - p_t)1[p_t \leq m_t] + c.$$

The feedback the vendor receives is

$$F_{p_t,m_t} = 1[p_t \leq m_t].$$

Discrete decision problems with partial monitoring

Example: **Label efficient prediction**

Aim is to predict a sequence of outcomes ($y_t \in \{1, \ldots, k\}$). At round t:

1. A prediction strategy either predicts $\hat{y}_t$ and incurs a loss $L_{\hat{y}_t, y_t} = 1[\hat{y}_t \neq y_t]$ (but the outcome y_t is not revealed), or
2. The strategy buys the label ($\hat{y}_t = 0$), incurs loss $c \in [0, 1]$, and the outcome y_t is revealed.

$$F_{\hat{y}_t, y_t} = 1[\hat{y}_t = 0] y_t.$$

Discrete decision problems with partial monitoring

Sequential decision problem. At each step:

1. strategy chooses (distribution of) $I_t \in \{1, \ldots, k\}$ and environment chooses (distribution of) $J_t \in \{1, \ldots, m\}$.
2. strategy incurs loss L_{I_t, J_t} (but does not see it).
3. strategy receives feedback F_{I_t, J_t}.

Two flavors:

Stochastic The environment can choose J_t i.i.d.

Adversarial The environment chooses J_t with full knowledge of all previous choices.

Discrete decision problems with partial monitoring

- The loss matrix $L \in \mathbb{R}^{k \times m}$ and feedback matrix $F \in \mathbb{N}^{k \times m}$ are fixed and known.
- The aim of the strategy is to minimize regret,

$$R_n = \sum_{t=1}^{n} L_{I_t, J_t} - \min_i \sum_{t=1}^{n} L_{i, J_t},$$

(in expectation or with high probability) or pseudo-regret,

$$\overline{R}_n = \mathbb{E} \sum_{t=1}^{n} L_{I_t, J_t} - \min_i \mathbb{E} \sum_{t=1}^{n} L_{i, J_t}.$$

Partial monitoring: Examples

Dynamic pricing:

$$L_{p_t,m_t} = (m_t - p_t)1[p_t \leq m_t] + c,$$
$$F_{p_t,m_t} = 1[p_t \leq m_t].$$

Dynamic pricing variant:

$$L_{p_t,m_t} = c - p_t 1[p_t \leq m_t],$$
$$F_{p_t,m_t} = 1[p_t \leq m_t].$$

(bandit!)

Partial monitoring: Examples

Label efficient prediction:

$$L_{\hat{y}_t, y_t} = 1[\hat{y}_t \neq y_t],$$
$$F_{\hat{y}_t, y_t} = 1[\hat{y}_t = 0] y_t.$$

General bandit problem:

$$L_{:,j} = \text{losses for outcome } j,$$
$$F_{i,j} = L_{i,j}.$$

Full information: $F_{i,j} = j$.

Partial monitoring: Regret

For full information problems, the stochastic minimax regret and the adversarial regret are $\tilde{\Theta}(\sqrt{n})$.

For bandit problems, these are also $\tilde{\Theta}(\sqrt{n})$. (But with worse dependence on the number of arms.)

What is achievable for other partial monitoring problems?

Partial monitoring: Regret

Example:

$$L = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 2 & 1 \end{pmatrix} \qquad F = \begin{pmatrix} 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 3 \end{pmatrix}$$

Regret is 0: never need to try the first action.

Example:

$$L = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \qquad F = \begin{pmatrix} 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 3 \end{pmatrix}$$

Regret is $\Omega(n)$: adversary's choice between last two actions is always hidden.

Partial monitoring: Regret

Stochastic minimax regret:

0	if trivial (only one nondegenerate action)
$n^{1/2}$	if nontrivial and locally observable
$n^{2/3}$	if observable but not locally observable
n	if not observable.

(ignoring log factors).

Partial monitoring: Regret

(Oblivious) adversarial regret:

0	if trivial
$\Omega(n^{1/2})$	if nontrivial
$O(n^{1/2})$	if nontrivial, NDD, locally observable
$\Omega(n^{2/3})$	if observable, NDD, not locally observable
$O(n^{2/3})$	if observable
n	if not observable.

(ignoring log factors). 'NDD' means 'no degenerate or duplicate actions.'

Partial monitoring: Definitions

Define L^i as the ith row of L.

Definition: The *optimal cell* for action i is the subset of the m-simplex on which it gives the least expected loss:

$$N_i = \left\{ p \in \Delta_m : \forall j,\, L^i \neq L^j \Rightarrow L^i p \leq L^j p \right\}.$$

Action i is *dominated* if N_i is empty.

Clearly, if all but one action is dominated, it suffices to play that action to get zero regret. But we can also avoid playing actions that are almost dominated...

Partial monitoring: Degenerate actions

Lemma: If we define the *open optimal cell* for action i as

$$S_i = \left\{ p \in \Delta_m : \forall j,\, L^i \neq L^j \Rightarrow L^i p < L^j p \right\},$$

then for any $p \in \Delta_m$, there is an i with S_i non-empty such that $p \in N_i$.

Hence, we don't need to worry about exploiting *degenerate* actions (but we might need to use them to distinguish losses of other actions).

Definition: An action i is *degenerate* if S_i is empty.

Partial monitoring: Regret lower bounds

Why does nontrivial imply $\Omega(\sqrt{n})$?

When the problem is nontrivial, there is a boundary between optimal cells. By choosing a distribution for the adversary (randomly) that is ϵ to one side or the other of that boundary, the regret of mistaking the optimal action is of order ϵn. But even if the adversary's actions are observed, the fluctuations in their relative frequency will scale like $1/\sqrt{n}$ (and things certainly cannot be improved by seeing limited feedback). Choosing ϵ of this scale will give the $\sqrt{n}$ lower bound.

Partial monitoring: $O(n^{2/3})$ regret

To illustrate the idea, we'll look first at a weaker result. Suppose we can write $L = KF$ for some matrix $K \in \mathbb{R}^{k \times k}$. Then

$$L_{i,j} = \sum_l K_{i,l} F_{l,j},$$

so we can use

$$\tilde{L}_{i,J_t} = \frac{K_{i,I_t} F_{I_t,J_t}}{p_{I_t,t}}$$

as an unbiased estimate of the loss L_{i,J_t}, and it only needs to see the feedback F_{I_t,J_t}.

Partial monitoring: $O(n^{2/3})$ regret

The idea is to use this estimate with an exponential weights strategy, where the exponential distribution is mixed with a uniform distribution over actions (the mixture component decreases slowly, as $t^{-1/3}$; this constrains $1/p_{I_t,t}$, and can be viewed as an exploration component—we'll see why it's essential in general).

Theorem: For $n = \tilde{\Omega}(k^2 \log^3 1/\delta)$, with probability at least $1-\delta$,

$$R_n \leq ck^{2/3} (\log k)^{1/3} n^{2/3} \sqrt{\log 1/\delta}.$$

Partial monitoring: $O(n^{2/3})$ regret

- Constants involve size of entries of K matrix.
- Notice poor dependence on k. If there is a revealing action (one that reveals action of adversary), then by playing it randomly (roughly a proportion $n^{-1/3}$ of the time) and using the revealed adversary action to estimate the cumulative losses, it is possible to obtain high probability regret bounds that grow as

$$n^{2/3} \log^{1/3}(k/\delta).$$

- How to extend beyond $L = KF$?

Partial monitoring: Regret

Recall L^i is the ith row of L and define F^i as the ith row of F.

Definition: For each $i \in \{1, \ldots, k\}$ if the entries of F^i are $f_1, \ldots, f_{m_i}$, define the *signal matrix* $S^i \in \mathbb{R}^{m_i \times m}$ as

$$S^i_{j,l} = 1[F^i_l = f_j].$$

(L, F) is *observable* if, for all nondegenerate i, j,

$$L^i - L^j \in \text{span}\left(\bigcup_k \{\text{rows of } S^k\}\right).$$

The $L = KF$ idea extends in this case to estimating the nondegenerate rows of L using the feedback. This approach gives $n^{2/3}$ regret bound for the observable case.

Partial monitoring: Regret lower bounds

If (L, F) is not observable, the regret is $\Omega(n)$.

High level idea:
There are two nondegenerate actions, say i and j, whose average losses cannot be distinguished via observing feedback. Then can construct two different values of the adversary's distribution p, one in S_i and one in S_j, that lie in the subspace orthogonal to the observed space

$$\text{span}\left(\bigcup_k \{\text{rows of } S^k\}\right),$$

but have a non-zero inner product with $L^i - L^j$ (that is, the expected losses differ). Then the distinction between these two adversarial probability distributions will never be observed. So the expected regret (under a random choice of those two) will grow linearly.

Partial monitoring: Local observability

We are concerned with distinguishing adversary distributions near a boundary between cells of nondegenerate actions. To get good performance while distinguishing between these distributions, we must be able to estimate differences of losses using actions that are optimal at the boundary.

Definition: Actions i, j are locally observable if they are non-degenerate, their optimal cells share a boundary that is $(m-2)$-dimensional, and

$$L^i - L^j \in \text{span}\left(\bigcup\{\text{rows of } S^k : N_i \cap N_j \subset N_k\}\right).$$

Partial monitoring: Local observability

It turns out that without local observability, the regret in the stochastic (hence adversarial) setting grows as $\Omega(n^{1/3})$: distinguishing the losses of two nondegenerate actions that are not locally observable requires an action that is far from optimal. This requires a separation of exploration and exploitation, which leads to the $n^{2/3}$ regret:

Suppose that we explore a proportion γ of the time, incurring a constant regret for each exploration trial, and exploit the remaining time, incurring a regret per trial that decreases no faster than $(\gamma n)^{-1/2}$. Then regret will scale like

$$\gamma n + \frac{(1-\gamma)n}{(\gamma n)^{1/2}},$$

which is minimized for $\gamma \sim n^{-1/3}$, giving regret of order $n^{2/3}$.

Partial monitoring: Local observability

With local observability in the stochastic setting, an upper confidence bound strategy can be constructed that works separately for each 'local pair.' In the (oblivious) adversarial setting, it suffices to ensure a bound on 'local' internal regret (internal regret means the decrease in cumulative loss that would have resulted from consistently substituting one action for another; local means only substituting neighboring actions).