Online Learning with Memory and Non-stochastic Control

Peng Zhao, Yu-Xiang Wang, Zhi-Hua Zhou

Contact

{zhaop, zhouzh}@lamda.nju.edu.cn

yuxiangw@cs.ucsb.edu

Online Learning to Online Decision Making

Standard Online Convex Optimization:

The loss of the t-th round is only related to the decision $\mathbf{w}_t$

Goal: to predict as well as the best offline decision

Online Convex Optimization with Memory

The loss of the t-th round can depend on the *historical decisions*, for example, related to the past $m+1$ decision $\mathbf{w}_t, \mathbf{w}_{t-1}, \dots, \mathbf{w}_{t-m}$

⇨ *a simplified model to capture the memory effect in online decision making*

Policy regret: $\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \min_{\mathbf{v}\in\mathcal{W}} \sum_{t=1}^{T} f_t(\mathbf{v}, \dots, \mathbf{v})$

Non-stationary Environments: online learning for real-world applications (such as whether forecasting, electricity prediction, etc)

→ optimal decision usually *changes* in non-stationary environments

Non-stationary OCO with Memory

Dynamic Policy Regret: competing with ***any*** comparators $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_T$

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) = \sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \sum_{t=1}^{T} f_t(\mathbf{v}_{t-m:t})$$

adaptive to non-stationarity of environments

universal guarantee against any comparator sequence

specialize → *Static* policy regret of OCO w memory, when $\mathbf{v}_1 = \cdots = \mathbf{v}_T = \mathbf{v}^*$

$$\text{S-Regret}_T(\mathbf{v}^*) = \sum_{t=1}^{T} f_t(\mathbf{w}_{t-m:t}) - \sum_{t=1}^{T} f_t(\mathbf{v}^*, \dots, \mathbf{v}^*)$$

specialize → *Dynamic* regret of standard OCO, when memory length $m = 0$

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{v}_t)$$

Dynamic Regret Optimization

Reduction to OCO with switching cost: by exploiting the coordinate-wise Lipschitzness of the loss function f_t, the dynamic regret of OCO with memory can be upper bounded by three parts.

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) \le \underbrace{\sum_{t=1}^{T} \tilde{f}_t(\mathbf{w}_t) - \sum_{t=1}^{T} \tilde{f}_t(\mathbf{v}_t)}_{\text{unary regret on } \tilde{f}_{1:T}} + \lambda \underbrace{\sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2}_{\text{switching cost}} + +\lambda \underbrace{\sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2}_{\text{path-length}}$$

where $\tilde{f}_t(\mathbf{w}) \coloneqq f_t(\mathbf{w}, \dots, \mathbf{w})$

Algorithm: run Online Gradient Descent (OGD) over the unary loss function $\tilde{f}_1, \dots, \tilde{f}_T$

→ *as it naturally optimizes unary regret, also moves slow enough to minimize switching cost*

Regret : when path-length $P_T = \sum_{t=2}^{T} \|\mathbf{v}_t - \mathbf{v}_{t-1}\|_2$ is known in advance, it gives an optimal $\mathcal{O}(\sqrt{T(1+P_T)})$ dynamic regret

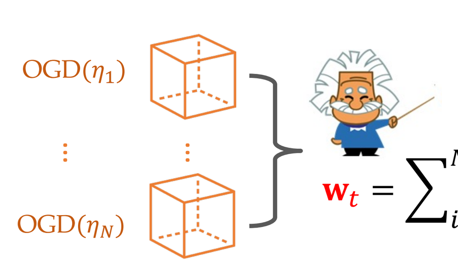
but, environmental non-stationarity is unknown

Challenge: Unknown Path-length and Switching Cost

Unknown Path-length P_T:

The optimal step size is $\mathcal{O}(\sqrt{(1+P_T)/T})$, hence requiring knowledge of path-length P_T as the algorithmic input (which is unfortunately *unknown*).

Online Ensemble with meta-base aggregation: run multiple base-learners with different step sizes to hedge non-stationarity, and employ a meta-learner for adaptive ensemble



$\mathbf{w}_t = \sum_{i=1}^{N} p_{t,i}\, \mathbf{w}_{t,i}$

Step size pool:

$$\mathcal{H} = \left\{ \eta_i \,\middle|\, \eta_i = 2^{i-1} \cdot \sqrt{\frac{D^2}{(\lambda G + G^2)T}}, i \in [N] \right\}$$

$$\text{D-Regret}_T(\mathbf{v}_{1:T}) \le \mathcal{O}\left(\sqrt{T(1+P_T)}\right) + \mathcal{O}(P_T) + \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2$$

Switching Cost (key entity in OCO with memory):

Tension between dynamic regret and switching cost: optimizing dynamic regret requires the algorithm to *move fast* to catch up with the environment, which is contradictory with *low switching cost*.

$$\sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2 \le D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1 + \sum_{t=2}^{T} \sum_{i=1}^{N} p_{t,i} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$$

maximum step size: $\eta_N = \mathcal{O}(1)$ — *grows linearly in T!*

switching cost of this learner: $\|\mathbf{w}_{t,N} - \mathbf{w}_{t-1,N}\|_2 \le \mathcal{O}(\eta_N T) = \mathcal{O}(T)$!

Solution: Algorithmically Enforce Low Switching Cost

Main Idea: Algorithmically Enforce Low Switching Cost

Avoid directly controlling the switching cost but adding it as a penalty term into the loss function. Use algorithm to enforce low switching cost.

♠ **Technical contributions**: a switching-cost-regularized surrogate loss

By introducing a novel *switching-cost-regularized surrogate loss*, we obtain a novel meta-base regret decomposition.

Dynamic Regret Decomposition (suppose we run N base learners):

meta regret: $\sum_{t=1}^{T} (\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \ell_{t,i}) + \lambda D \sum_{t=2}^{T} \|\boldsymbol{p}_t - \boldsymbol{p}_{t-1}\|_1$

base regret: $\sum_{t=1}^{T} \langle \nabla \tilde{f}_t(\mathbf{w}_t), \mathbf{w}_{t,i} - \mathbf{v}_t \rangle + \lambda \sum_{t=2}^{T} \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$

where $\mathbf{w}_t$ is the final decision sequence as output;

$\mathbf{w}_{t,i}$ is the prediction sequence of the i-th base learner, for any $i \in [N]$;

$$\ell_{t,i} \coloneqq \langle \nabla \tilde{f}_t(\mathbf{w}_t), \mathbf{w}_{t,i} \rangle + \lambda \|\mathbf{w}_{t,i} - \mathbf{w}_{t-1,i}\|_2$$

is the switching-cost-regularized surrogate loss of the meta learner.

Algorithm: Switching-Cost-Regularized Ensemble Algorithm for OCO with Memory ***(Scream)***

meta learner: Hedge, which updates as $p_{t+1,i} \propto p_{t,i} \exp(-\epsilon \ell_{t,i})$;

base learner: Online Gradient Descent, which updates as $\mathbf{w}_{t+1,i} = \Pi_{\mathcal{W}}[\mathbf{w}_{t,i} - \eta_i \nabla \tilde{f}_t(\mathbf{w}_t)]$.

Regret: Scream enjoys a *minimax optimal* $\mathcal{O}\left(\sqrt{T(1+P_T)}\right)$ dynamic policy regret.

Application: Online Non-stochastic Control

Linear Dynamical System: $x_{t+1} = Ax_t + Bu_t + w_t$

Online Non-stochastic Control:

At each round $t = 1, 2, \dots, T$

1. the player observes a state x_t and provides a control u_t;
2. the player suffers a convex loss $c_t(x_t, u_t)$;
3. the environment chooses an *adversarial* noise w_t and evolves to state x_{t+1}.

Dynamic Policy Regret : competing with any controllers $\pi_1, \pi_2, \dots, \pi_T$

$$\text{D-Regret}_T(\pi_{1:T}) = \sum_{t=1}^{T} c_t(x_t, u_t) - \sum_{t=1}^{T} c_t(x_t^{\pi_t}, u_t^{\pi_t})$$

Reduction to OCO with Memory:

Policy parametrization: Disturbance-Action Controller (DAC)

$u_t = -Kx_t + \sum_{i=1}^{H} M^{[i]} w_{t-i}$ where K and M are controller parameters with certain assumptions.

Truncation: under mild conditions, the states and actions that are more than m rounds before can be truncated at an acceptable cost

$$\text{D-Regret}_T(M_{1:T}^*) = \sum_{t=1}^{T} f_t(M_{t-m:t}) - \sum_{t=1}^{T} f_t(M_{t-m:t}^*)$$

where $M_{1:T}$ are the parameters of our controller while $M_{1:T}^*$ are those of the comparator controllers. Note that $f_{1:T}$ are truncated losses sent to the OCO with memory.

Result: the *first* controller with *dynamic policy regret* for non-stochastic control, by employing Scream to optimize the reduced OCO with memory, with an $\tilde{\mathcal{O}}\left(\sqrt{T(1+P_T)}\right)$ dynamic policy regret, where $P_T = \sum_{t=2}^{T} \|M_t - M_{t-1}\|_F$.