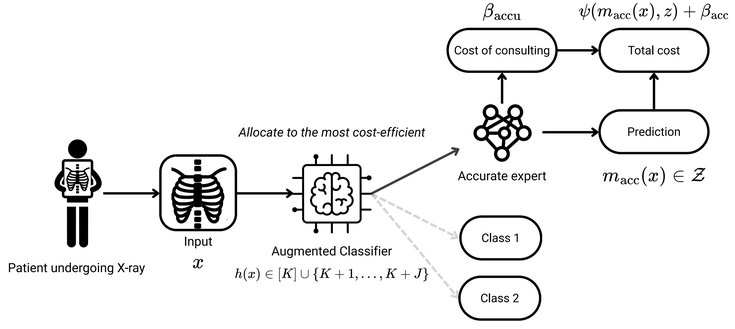
Online Learning-to-Defer with Varying Experts

Dang Hoang Duy[*1] Yannis Montreuil[*1,4,5] Maxime Meyer[*4,6] Axel Carlier[2,4] Lai Xing Ng[3,4] Wei Tsang Ooi[1,4]

[1]School of Computing, National University of Singapore [2]Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse [3]Institute for Infocomm Research, A*STAR
[4]IPAL, IRL 2955, Singapore [5]CNRS@CREATE LTD, Singapore [6]Department of Mathematics, National University of Singapore

Problem

Learning-to-Defer (L2D) routes each query either to a model or to an expert. In the **one-stage** setting, the predictor and the allocation policy are learned *jointly*.



One-Stage L2D Setup

Inputs $x \in \mathcal{X}$, labels $y \in \{1, \ldots, K\}$, and J experts with outputs $\mathbf{m} = (m_1, \ldots, m_J)$.
A hypothesis $h \in \mathcal{H}$ scores the **augmented action space** $\mathcal{A} = \{1, \ldots, K+J\}$: the first K actions are class predictions, the last J are deferrals. The true deferral loss is

$$\ell_{\text{def}}(\hat{h}(x), y, \mathbf{m}) = \begin{cases} \mathbf{1}\{\hat{h}(x) \neq y\}, & \hat{h}(x) \leq K, \\ c_j(m_j, y), & \hat{h}(x) = K+j, \end{cases}$$

with expert cost $c_j(m_j, y) = \mathbf{1}\{m_j \neq y\} + \beta_j$ (β_j: consultation cost).

Motivation

We introduce the Online setting. This takes into account experts that may evolve over time (e.g. availability, accuracy, area of expertise). At every round t:

- the learner commits to a linear strategy h_t,
- he receives an adversarially chosen input $\overline{x}_t = (x_t, A_t)$.
- A true label $y_t \sim p(\cdot \mid \overline{x}_t)$ is chosen stochastically, but remains hidden to the learner.
- The learner incurs a loss $\ell_{\text{def}}(h_t, \overline{x}_t, y_t)$.

Goal. Minimize the cumulative loss compared to the best strategy in hindsight.

Contributions. Our method achieves regret guarantees of $O((n+n_e)T^{2/3})$ in general and $O((n+n_e)\sqrt{T})$ under a low-noise condition, where T is the time horizon, n is the number of labels, and n_e is the number of distinct experts observed across rounds.

$\mathcal{H}$-Consistency Bound for Online Learning

A central property in batch learning is consistency, which ensures that minimizing a surrogate risk recovers the Bayes-optimal solution.

Definition 1 (Surrogate Deferral Loss). For any $h \in \mathcal{H}$ and $(\overline{x}, y) = ((x, A), y) \in \overline{\mathcal{X}} \times \mathcal{Y}$,

$$\Phi_{\text{def}}(h, \overline{x}, y) = \widetilde{\Phi}_{01}(h, \overline{x}, y) + \sum_{n+j \in \overline{\mathcal{Y}}_A} (1 - c_j(x, y)) \widetilde{\Phi}_{01}(h, \overline{x}, n+j).$$

Theorem 1. Assume that $\widetilde{\Phi}_{01}$ admits an $\mathcal{H}$-consistency bound w.r.t ℓ_{01} with a concave, non-decreasing calibration function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ with $\Gamma(0) = 0$. Then for any $\overline{\mathbf{x}} \in \overline{\mathcal{X}}^T$ and any distribution $p \in \mathcal{D}$,

$$\frac{R_{\ell_{\text{def}}}(T) + \mathcal{M}_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}}) - c\left(\sum_{t \in [T]} \gamma_t\right)}{T} - \epsilon \leq \max_{t \in [T]} S(\overline{x}_t)\, \Gamma\left(\frac{R_{\Phi_{\text{def}}}(T) + \mathcal{M}_{\Phi_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}})}{T \min_{t \in [T]} S(\overline{x}_t)}\right).$$

The regret typically depends on the $(\ell_{\text{def}}, \mathcal{H})$-minimizability gap $\mathcal{M}_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}}) = \mathcal{R}^*_{\ell_{\text{def}}, \mathcal{H}}(\overline{\mathbf{x}}) - \sum_{t \in [T]} \mathcal{C}^*_{\ell_{\text{def}}, \mathcal{H}}(\overline{x}_t)$. The minimizability gap measures the discrepancy between selecting the best hypothesis adaptively at each round and selecting the best single hypothesis in hindsight.

Regret bounds for Online L2D

When the surrogate is chosen as the constrained hinge loss, the task can be reduced to a constrained online convex optimization problem.

OGD with Projected Gradients

Input: Learning rate $\eta_t > 0$, and exploration rate γ_t
Initialize: $\widetilde{W}_1 = \mathbf{0} \in \mathbb{R}^{N \times (d+1)}$
for $t = 1$ **to** T **do**
  Obtain $\overline{x}_t = (x_t, A_t)$
  Project $\widetilde{W}_t$ to feasible hypothesis set: $\widetilde{W}_t|_{\mathcal{K}_{A_t}} = \Pi_{\mathcal{K}_{A_t}}(\widetilde{W}_t)$
  Let $y_t^* = \arg\max_y \langle w_{t,y}, x_t \rangle + b_t$
  Set $q_t = (1 - \gamma_t)\mathbf{e}_{y_t^*} + \frac{\gamma_t}{|\mathcal{Y}_{A_t}|}\mathbf{1}$
  Predict with label $y'_t \sim q_t$
  Obtain feedback and compute $\widehat{\Phi}_t(\widetilde{W}_t|_{\mathcal{K}_{A_t}})$ accordingly
  Compute gradient $\widehat{\nabla}_t = \nabla \widehat{\Phi}_t(\widetilde{W}_t|_{\mathcal{K}_{A_t}})$ and its projection onto $\mathcal{K}_{A_t}$: $\widehat{\nabla}_t|_{\mathcal{K}_{A_t}} = \Pi_{\mathcal{K}_{A_t}}(\widehat{\nabla}_t)$
  Update $\widetilde{W}_{t+1} = \widetilde{W}_t - \eta_t \widehat{\nabla}_t|_{\mathcal{K}_{A_t}}$
end for

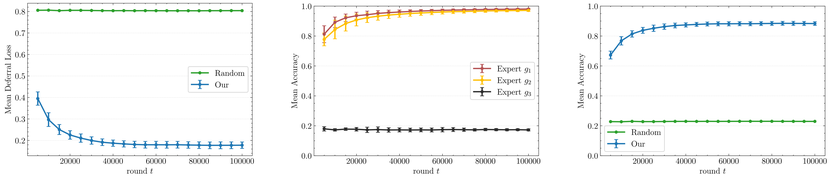
Theorem 2. Running OGD with Projected Gradients with $\eta_t = \frac{B}{\sqrt{N}Rt^{2/3}}$, $t = 1, \ldots, T$, $\gamma_t = \min\left\{\frac{1}{2}, \frac{1}{t^{1/3}}\right\}$ gives us the following expected regret bound

$$R_{\ell_{\text{def}}}(T) = \sum_{t \in [T]} \left(\ell_{\text{def}}(q_t, \overline{x}_t, y_t) - \ell_{\text{def}}(\widetilde{W}^*, \overline{x}_t, y_t)\right) = O\left(N^{3/2} BRT^{2/3}\right).$$

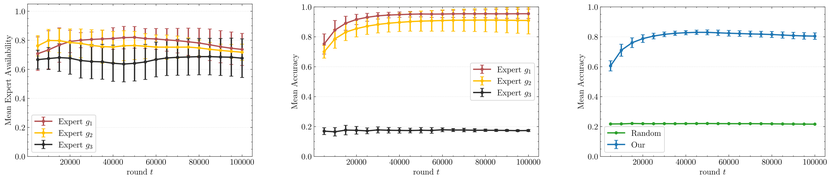
Experiments

Setup. We study three different datasets (Synthetic Datasets, Reuters4 and CIFAR-10H). We run our algorithm under different expert regimes and obtain similar results for all three datasets. We show our results on only one dataset for clarity.

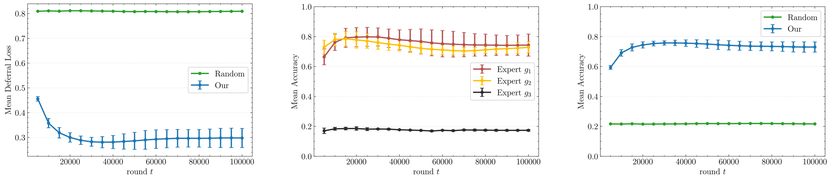
- Fixed expertise and availability.



- Fixed expertise but varying availability.



- Varying expertise and availability.



Takeaway

- We can generalize L2D methods to varying experts.
- This is achieved by deriving novel H-consistency bounds and applying online gradient descent.