An Online Learning Approach to Interpolation and Extrapolation in Domain Generalization

Elan Rosenfeld Pradeep Ravikumar Andrej Risteski

Machine Learning Department, Carnegie Mellon University

Introduction

Typical domain generalization analysis defines an ***uncertainty/perturbation set*** and solves for the minimax predictor.

- (Distributionally Robust Optimization, Group Shift Robustness, Invariant Causal Prediction, etc.)

But this is *needlessly pessimistic*;
The Universe isn't actively trying to make us fail!

Instead, we consider an **online game**.
Over T rounds, for each round *t:*

1. Player chooses predictor $f_t : \mathcal{X} \to \mathcal{Y}$.
2. Nature plays domain $p_t \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ (distribution over samples).
3. Player suffers loss $L_t(f_t) := \mathbb{E}_{p_t}[\ell(f_t(x), y)]$.

Player's goal is to ***minimize regret***

$$\sum_{t=1}^{T} L_t(f_t) - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} L_t(f) .$$

Captures our desire to improve and perform reasonably well on all distributions when environments **aren't necessarily adversarial.**

Interpolation vs Extrapolation

How do we define nature's playable region $\mathcal{P}$? Common to split uncertainty set into two "regimes": ***Interpolation*** and ***Extrapolation***. This is somewhat vague and informal.

It's also folklore that Empirical Risk Minimization (ERM) is able to interpolate but *cannot extrapolate*.

We analyze these two regimes according to an existing notion of domain generalization:
Linear combinations of domain likelihoods.

Interpolation:

Given domains $\mathcal{E} = \{e_1, \ldots, e_E\}$, playable region is all distributions that can be written as

$$p(x, y) = \sum_{e \in \mathcal{E}} \lambda^e p^e(x, y)$$

for some $\lambda \in \Delta_E$, the E-simplex (i.e., convex combinations).

Extrapolation:

Playable region is all **functions** (not always distributions!) that can be written as "bounded affine" combinations

$$p(x, y) = \sum_{e \in \mathcal{E}} \lambda^e p^e(x, y)$$

where

$$\sum \lambda^e = 1, \quad \lambda^e \geq -\beta \; \forall e .$$

What can we say about (info-theoretic) achievable regret in this setting?

Theorem 1: Under *interpolation*, regret is lower bounded as $\Omega(\log T)$. This regret is achieved by Follow-the-Leader (FTL). In our game, **FTL is exactly Empirical Risk Minimization** **(ERM is minimax optimal!)**

Theorem 2: Under *extrapolation,* regret is lower bounded as $\Omega(\sqrt{T})$. This regret is achieved by Follow-the-Perturbed-Leader (FTPL). **FTPL is a (noisy) version of ERM** **(ERM remains optimal for extrapolation!)**

Computational Complexity

Information-theoretically, the **minimax optimal regret is achievable by ERM** for both interpolation and extrapolation.

This is quite surprising, as conventional wisdom holds that extrapolation is *significantly* harder (statistically) than interpolation.
This implies linear combinations of domain likelihoods *may not be the correct model of domain generalization.*

What about **computationally**? What can we say about the computational difficulty of extrapolation?

Theorem 3: Under *extrapolation*, even against an *oblivious adversary* who must choose the full sequence of domains ahead of time, **achieving sublinear regret is NP-Hard.**

Thus domain extrapolation is *exponentially harder* than interpolation in this setting.