Adaptive Regularization

Improving Invariant Risk Minimization

Badr Youbi Idrissi

CentraleSupelec

25 January 2021

Introduction

- Astonishing results on benchmarks
- Learns complex statistical dependences
- What happens when underlying distribution shifts?

Figure 1: Illustration of Shortcut learning[1]

[1] Geirhos et al., *Shortcut Learning in Deep Neural Networks*.

Simple Linear Example

Let X, Y and Z be real random variables such that

$$X := \mathcal{N}(0, \sigma^2) \tag{1}$$

$$Y := X + \mathcal{N}(0, \sigma^2) \tag{2}$$

$$Z := Y + \mathcal{N}(0, 1) \tag{3}$$

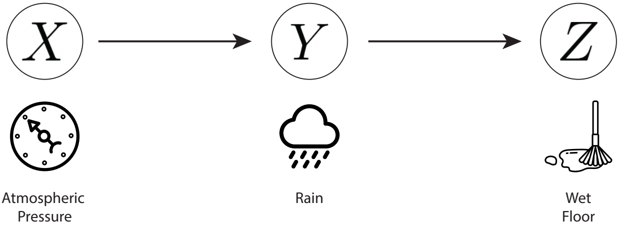


Figure 2: Simple Linear Example[2]

Regression Results

We have three cases :

- If we regress *Y* over *X* we would have $\hat{\alpha_1} = 1$ and $\hat{\alpha_2} = 0$
- If we regress *Y* over *Z* we would have $\hat{\alpha_1} = 0$ and $\hat{\alpha_2} = \frac{\sigma^2}{\sigma^2+1/2}$
- If we regress *Y* over *X* and *Z* we would have $\hat{\alpha_1} = \frac{1}{\sigma^2+1}$ and $\hat{\alpha_2} = \frac{\sigma^2}{\sigma^2+1}$

Only first regression doesn't depend on the standard deviation is the first.

$\Rightarrow$ Collecting data over different cities : detect invariance.

If decision depends on environment then we can make error arbitrarily big.

Invariant Risk Minimization

- Extract an invariant representation ?
- Normally minimize $R(f) = E[\text{loss}(f(X), Y)]$
- When access to multiple environments $R^e(f) = E[\text{loss}(f(X^e), Y^e)]$ with e an environment.
- Split $f = w \circ \Phi$. Φ is the representation. w is a "classifier".
- Φ is invariant if $\forall e \quad w^* = \arg\min_w R^e(w \circ \Phi)$

$$\begin{array}{ll} \min_{\Phi:\mathcal{X}\to\mathcal{H}} & \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w \circ \Phi) \\ \text{subject to} & w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min} R^e(\bar{w} \circ \Phi), \text{ for all } e \in \mathcal{E}_{\text{tr}}. \end{array}$$

Simplified IRMv1

- Complex 2 level optimization
- We fix the classifier to a linear one and let Φ make it optimal
- Since it is linear optimality can be measured with $\|\nabla_w R^e(w \circ \Phi)\|$

$$\min_{\Phi:\mathcal{X}\to\mathcal{H}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w \circ \Phi) + \lambda \|\nabla_w R^e(w \circ \Phi)\|$$

iBadr

Limitations and Proposal

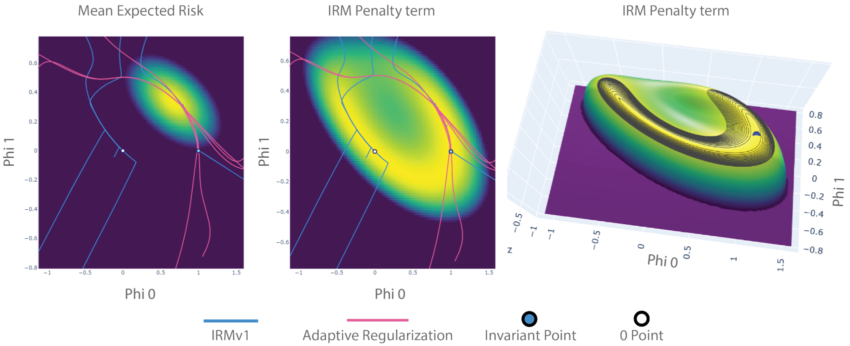
- In non linear case can find non invariant solution with small penalty
- Sensitive to initialization. We propose :

Algorithm 2: Adaptive Regularization

Input: $\lambda_{init}, T, M, \beta$
Output: Φ
Data: Training data

$\lambda := \lambda_{init} * \text{ones\_like}(\Phi)$
while *Current Iteration* $<$ *Max Iterations* **do**
  $g_e = \nabla_\Phi R^e(w \circ \Phi)$
  $g_p = \nabla_\Phi \|\nabla_w R^e(w \circ \Phi)\|$
  $g = g_e + \lambda * g_p$
  $e = \beta e + (1 - \beta) g^2$
  for i *such that* $e[i] < T$ **do**
    $\lambda[i] := M\lambda[i]$

Explanation



Results

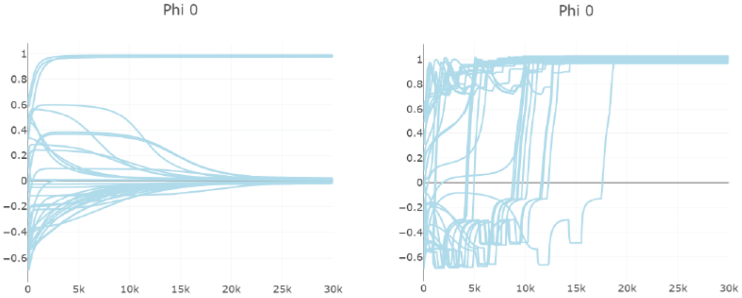


Figure 3: IRMv1 on the left. Adaptive Regularization on the right

Results

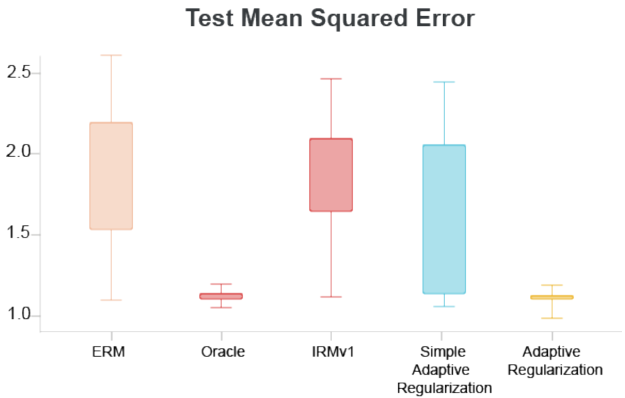


Figure 4: Test losses