Group Distributionally Robust Optimization (Group DRO)

Presented by Haoxiang Wang, Lang Yin and Haozhe Si
Comparison of robust optimization with distributionally robust optimization.

Deterministic Optimization
\[
\min_{\theta} \ell(\theta; (x, y))
\]

Stochastic Optimization
\[
\min_{\theta} \mathbb{E}_{x, y \sim P} \left[ \ell(\theta; (x, y)) \right]
\]

Robust Optimization
\[
\min_{\theta} \max_{\tilde{x} \in \Delta} \ell(\theta; (\tilde{x}, y))
\]

Distributionally Robust Optimization
\[
\min_{\theta} \max_{Q \in \mathcal{Q}} \mathbb{E}_{x, y \sim Q} \left[ \ell(\theta; (x, y)) \right]
\]

The graphic illustration is inspired: [Chen and Paschalidis]. Distributionally Robust Learning. arXiv:2108.08993
Empirical Risk Minimization (ERM):

\[ \theta^*_{\text{ERM}} = \min_{\theta \in \Theta} \mathbb{E}_{x, y \sim P} [\ell(\theta; (x, y))] \quad \text{finite samples} \]

\[ \hat{\theta}_{\text{ERM}} = \min_{\theta \in \Theta} \mathbb{E}_{x, y \sim \hat{P}} [\ell(\theta; (x, y))] \]

where \( P \) is the ground-truth data distribution and \( \hat{P} \) is empirical distribution (finite samples)

Distributionally Robust Optimization (DRO)

\[ \min_{\theta \in \Theta} \left\{ R(\theta) := \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x, y) \sim Q} [\ell(\theta; (x, y))] \right\} \iff \min_{\theta \in \Theta} \max_{Q \in \mathcal{Q}} \mathbb{E}_{(x, y) \sim Q} [\ell(\theta; (x, y))] \]

where \( Q \) is an uncertainty set of distributions
Group DRO

DRO

\[
\min_{\theta \in \Theta} \left\{ \mathcal{R}(\theta) := \sup_{Q \in Q} \mathbb{E}_{(x,y) \sim Q} [\ell(\theta; (x,y))] \right\}
\]

where \( Q \) is an uncertainty set of distributions

Group DRO

\[
\min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \hat{P}_g} [\ell(\theta; (x,y))] \right\}
\]

where each group \( \hat{P}_g \) is an empirical distribution over all training points \( (x, y, g') \) with \( g' = g \) (or equivalently, a subset of training examples drawn from \( \hat{P}_g \))
Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input $x$: bird image

ML model

label: bird type

waterbird vs landbird

Wah et al. (2011), Zhou et al. (2017)
Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

Wah et al. (2011), Zhou et al. (2017)
Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input $x$: bird image

spurious correlation: land background

ML model

prediction $\hat{y}$: landbird

true label $y$: waterbird

Wah et al. (2011), Zhou et al. (2017)
Models can latch onto spurious correlations

input $x$: face image

ML model

label: hair color

blonde hair vs dark hair

Liu et al. (2015)
Models can latch onto spurious correlations

input $x$: face image

ML model

spurious correlation: gender
prediction $\hat{y}$: dark hair
true label $y$: blonde hair

Liu et al. (2015)
Models can latch onto spurious correlations

Models perform well on average

Average error: 0.03

But models can have high worst-group error

<table>
<thead>
<tr>
<th>label: object</th>
<th>waterbird</th>
<th>landbird</th>
</tr>
</thead>
<tbody>
<tr>
<td>spurious attribute: background</td>
<td></td>
<td></td>
</tr>
<tr>
<td>water background</td>
<td>0.05</td>
<td>0.21</td>
</tr>
<tr>
<td>land background</td>
<td>0.40</td>
<td>0.004</td>
</tr>
</tbody>
</table>

worst-group error: 0.40

Goal: low worst-group error

- Relies on spurious correlation
- High worst-group error

- More robust to spurious correlation
- Low worst-group error

Our approach: minimize the worst-group loss

Standard (ERM): average loss

$$\mathcal{R}_{\text{ERM}}(w) = \mathbb{E}_{(x, y, g)}[\ell(w; (x, y))]$$

Group DRO: worst-group loss

$$\mathcal{R}_{\text{gDRO}}(w) = \max_{g' \in G} \mathbb{E}_{(x, y, g)}[\ell(w; (x, y)) | g = g']$$

Train: known groups for each example
Test: unknown groups

Optimization algorithm for Group DRO

Optimizer

Group weights: which are worst-case?

✓ Scalable
✓ Theoretical guarantees
✓ Similar # iterations to convergence as ERM

Model parameters: Update on weighted loss

Optimizer

✓ ✓ ✓
✓ X ✓

Model

Group weights: which are worst-case?

✓ ✓ ✓
✓ X ✓

Online Optimization for Group DRO

**Objective:** \( \min_{\theta \in \Theta} \sup_{q \in \Delta_m} \sum_{g=1}^{m} q_g \mathbb{E}_{(x,y) \sim P_g} \left[ \ell(\theta; (x,y)) \right] \)

```
Algorithm 1: Online optimization algorithm for group DRO

Input: Step sizes \( \eta_q, \eta_\theta; P_g \) for each \( g \in \mathcal{G} \)
Initialize \( \theta^{(0)} \) and \( q^{(0)} \)
for \( t = 1, \ldots, T \) do
  \( g \sim \text{Uniform}(1, \ldots, m) \)  // Choose a group \( g \) at random
  \( x, y \sim P_g \)  // Sample \( x, y \) from group \( g \)
  \( q' \leftarrow q_{(t-1)} \);
  \( q'_g \leftarrow q'_g \exp(\eta_q \ell(\theta^{(t-1)}; (x,y))) \)  // Update weights for group \( g \)
  \( q^{(t)} \leftarrow q'/\sum_g q'_g \)  // Renormalize \( q \)
  \( \theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta q_g^{(t)} \nabla \ell(\theta^{(t-1)}; (x,y)) \)  // Use \( q \) to update \( \theta \)
end
```

Dynamics Group Loss Reweighting: \( \mathcal{L}(\theta^{(t)}) = \sum_{g=1}^{m} q_g^{(t)} \mathbb{E}_{(x,y) \sim P_g} \left[ \ell(\theta^{(t)}; (x,y)) \right] \)

Scale DOWN weight for majority group (easier to fit \( \rightarrow \) lower loss)
Scale UP weight for minority group (harder to fit \( \rightarrow \) larger loss)
Attempt 1: ERM $\rightarrow$ high worst-group test error

![Graph showing ERM error comparison between average and worst-group test.

Attempt 1: ERM $\rightarrow$ high worst-group test error

worst-group error is high because of poor generalization

Attempt 1: poor generalization → group DRO fails

worst-group error is high because of poor generalization but group DRO only controls *training* error!
New challenge: train error $\neq$ test error on worst group

Prior work: train error $\approx$ test error for worst-case group
• Small convex or generative models

Our setting: high worst-group test error despite zero train error
• State-of-the-art neural networks

Approach: regularization + group DRO

Problem: zero train error, but high worst-group test error

Solution: regularization

Counterintuitive with respect to recent trends:
More complex models with zero training error $\rightarrow$ better average error

Hoffer, Hubara, Soudry (2017), Belkin et al. (2019), Nakkiran et al. (2020)
Attempt 2: regularization + group DRO works

ERM with L2 penalty

Group DRO with L2 penalty

Group DRO + regularization mitigates the spurious correlation problem

Goal: low worst-group error

# Datasets

## Datasets

<table>
<thead>
<tr>
<th>Waterbirds</th>
<th>CelebA</th>
<th>MultiNLI</th>
</tr>
</thead>
</table>
| **y:** waterbird  
  **a:** water background | **y:** blond hair  
  **a:** female | **y:** contradiction  
  **a:** has negation |
| ![Waterbird Image] | ![Female Image] | (P) The economy could be still better.  
(H) The economy has never been better. |
| ![Landbird Image] | ![Male Image] | **y:** entailment  
  **a:** no negation |
| ![Waterbird Image] | ![Male Image] | (P) Read for Slate's take on Jackson's findings.  
(H) Slate had an opinion on Jackson's findings. |
| ![Landbird Image] | ![Male Image] | **y:** entailment  
  **a:** has negation |
| ![Landbird Image] | ![Male Image] | (P) There was silence for a moment.  
(H) There was a short period of time where no one spoke. |
## Average and Worst Case Accuracy for ERM and DRO

<table>
<thead>
<tr>
<th></th>
<th>Standard Regularization</th>
<th>Strong $\ell_2$ Penalty</th>
<th>Early Stopping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Accuracy</td>
<td></td>
<td>Worst-Group Accuracy</td>
</tr>
<tr>
<td></td>
<td>ERM</td>
<td>DRO</td>
<td>ERM</td>
</tr>
<tr>
<td>Waterbirds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Test</td>
<td>97.3</td>
<td>97.4</td>
<td>60.0</td>
</tr>
<tr>
<td>CelebA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>100.0</td>
<td>100.0</td>
<td>99.9</td>
</tr>
<tr>
<td>Test</td>
<td>94.8</td>
<td>94.7</td>
<td>41.1</td>
</tr>
<tr>
<td>MultiNLI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>99.9</td>
<td>99.3</td>
<td>99.9</td>
</tr>
<tr>
<td>Test</td>
<td>82.5</td>
<td>82.0</td>
<td>65.7</td>
</tr>
</tbody>
</table>
Connection to Multi-Task Learning

Takeaway

1. Spurious correlation $\rightarrow$ Low worst-group accuracy

2. Group Distributional Robust Optimization $\rightarrow$ (Dynamically) adjusts group weights to focus equally on majority & minority groups.
   - Related to Multi-Task Learning [1]

3. Regularization is necessary to Group DRO
   - Since model is over-parameterized (i.e., has large capacity) [2]

Thanks for watching this presentation!