# Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-case Generalization

CS 282R: Advancements in Probabilistic Machine Learning, ML Applications in Science, and Causality

Presented by Bahareh Tolooshams (I am not the author of this paper)

February 25, 2022

Harvard John A. Paulson
School of Engineering
and Applied Sciences

This presentation covers the following paper. I, Bahareh Tolooshams, am only the presenter of this work (not author).

This is part of CS 282R course, where we discussed papers, at Harvard University.

# Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization

**Shiori Sagawa***
Stanford University
ssagawa@cs.stanford.edu

**Pang Wei Koh***
Stanford University
pangwei@cs.stanford.edu

**Tatsunori B. Hashimoto**
Microsoft
tahashim@microsoft.com

**Percy Liang**
Stanford University
pliang@cs.stanford.edu

# Big Picture

## Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization

**Shiori Sagawa***
Stanford University
ssagawa@cs.stanford.edu

**Pang Wei Koh***
Stanford University
pangwei@cs.stanford.edu

**Tatsunori B. Hashimoto**
Microsoft
tahashim@microsoft.com

**Percy Liang**
Stanford University
pliang@cs.stanford.edu

Problem:

- Overparameterized networks can have high accuracy on average on in-domain test data, but fail on atypical examples.

How to solve:

- Use distributionally robust optimization (DRO) to account for worst-case training loss over certain groups.

# Spurious Correlations

One reason for failure:

- The network learns spurious correlations that hold on average but not in certain examples.

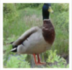
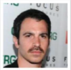
| | Common training examples | | Test examples |
|---|---|---|---|
| **Waterbirds** | **y:** waterbird<br>**a:** water background  | **y:** landbird<br>**a:** land background  | **y:** waterbird<br>**a:** land background  |
| **CelebA** | **y:** blond hair<br>**a:** female  | **y:** dark hair<br>**a:** male  | **y:** blond hair<br>**a:** male  |
| **MultiNLI** | **y:** contradiction<br>**a:** has negation<br>(P) The economy could be still better.<br>(H) The economy has never been better. | **y:** entailment<br>**a:** no negation<br>(P) Read for Slate's take on Jackson's findings.<br>(H) Slate had an opinion on Jackson's findings. | **y:** entailment<br>**a:** has negation<br>(P) There was silence for a moment.<br>(H) There was a short period of time where no one spoke. |

Figure 1: Representative training and test examples for the datasets we consider. The correlation between the label $y$ and the spurious attribute $a$ at training time does not hold at test time.

# Stochastic Optimization

- Input features $x \in \mathcal{X}$
- Predicting labels: $y \in \mathcal{Y}$

- Training $(x, y)$ are drawn from distribution $P$
- The empirical distribution is $\hat{P}$

Empirical risk minimization (ERM):

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

Generalization error:

$$|\mathbb{E}_{(x,y)\sim\hat{P}}[\ell(\theta; (x, y))] - \mathbb{E}_{(x,y)\sim P}[\ell(\theta; (x, y))]| \leq \epsilon$$

perform well on average on unseen data.

# Distributionally Robust Optimization (DRO)

Minimize the worst-case expected loss over an uncertainty set of distributions.

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

- $\hat{P} \in \mathcal{Q}$, i.e., divergence ball around the training distribution.
- Small ball: a regularizer.
- Large ball: a pessimistic approach on how well you know the true distribution.

# Group DRO (this paper)

Give some structure to the uncertainty sets using prior knowledge.

$$\mathcal{Q} := \left\{ \sum_{g=1}^{m} q_g P_g : q \in \delta_m \right\}$$

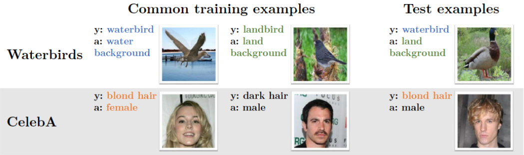
i.e., a mixture of groups constructed based on spurious correlations.

Minimize the worst-case loss over groups in the training data.

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

# How to Construct the Groups?

Make groups based on attributes that are spuriously correlated with the label.




Dataset:

- attributes: {male, female}, label: {blond, dark}

Groups

- $P_1$:{male, blond}, $P_2$:{male, dark}, $P_3$:{female, blond}, $P_4$:{female, dark}

**What spurious correlations the network learns in this case?**

**Between female and blond.**

So, the network do poorly on $P_1$: {male, blond}.

$$\hat{\theta}_{\text{DRO}} := \arg\min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \mathbb{E}_{(x,y)\sim \hat{P}_1} [\ell(\theta; (x,y))] \right\}$$

# How Good is this Grouping Approach?

- What if the relation between input and label is not as clear/simple as {gender attribute, hair color}?

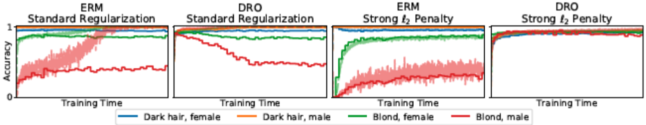
  **Example 2.1.** We want to automatically classify the quality of product reviews. Each review has a number of "helpful" votes $Y$ (from site users). We predict $Y$ using the text of the product review $X$. However, we find interventions on the sentiment $Z$ of the text change our prediction; changing "Great shoes!" to "Bad shoes!" changes the prediction.

- This grouping completely ignores the causal relationship between the input and label.

- Should we consider a more general construction where the causal relationships stay fixed, and the spurious dependencies change? (this is the topic of the other paper we discuss today).

# Group DRO vs. ERM

| | | | Average Accuracy | | Worst-Group Accuracy | |
|---|---|---|---|---|---|---|
| | | | ERM | DRO | ERM | DRO |
| Standard Regularization | Waterbirds | Train | 100.0 | 100.0 | 100.0 | 100.0 |
| | | Test | 97.3 | 97.4 | 60.0 | 76.9 |
| | CelebA | Train | 100.0 | 100.0 | 99.9 | 100.0 |
| | | Test | 94.8 | 94.7 | 41.1 | 41.1 |
| | MultiNLI | Train | 99.9 | 99.3 | 99.9 | 99.0 |
| | | Test | 82.5 | 82.0 | 65.7 | 66.4 |
| Strong $\ell_2$ Penalty | Waterbirds | Train | 97.6 | 99.1 | 35.7 | 97.5 |
| | | Test | 95.7 | 96.6 | 21.3 | 84.6 |
| | CelebA | Train | 95.7 | 95.0 | 40.4 | 93.4 |
| | | Test | 95.8 | 93.5 | 37.8 | 86.7 |
| Early Stopping | Waterbirds | Train | 86.2 | 80.1 | 7.1 | 74.2 |
| | | Test | 93.8 | 93.2 | 6.7 | 86.0 |
| | CelebA | Train | 91.3 | 87.5 | 14.2 | 85.1 |
| | | Test | 94.6 | 91.8 | 25.0 | 88.3 |
| | MultiNLI | Train | 91.5 | 86.1 | 78.6 | 83.3 |
| | | Test | 82.8 | 81.4 | 66.0 | 77.7 |



# Overparameterized DRO

DRO traditionally is applied when the training loss does not go to zero.

In overparameterized regime, training loss vanishes.

**What might be the problem here?**

- The network is optimal for both worst-case and average objectives.
- Good generalization on average, but bad generalization on worst-group.

**Strong regularization to avoid vanishing training loss regime.**

- $\ell_2$ penalty
- Early stopping
- Group adjustments

# General Idea behind Regularization

Constrain the model family's capacity to fit the training data.

Overparameterized network comes with implicit regularization.

- Norm regularizing of the weights through gradient descent.
- Good generalization on average, but not on worse-case.

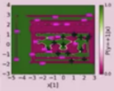
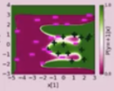
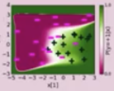
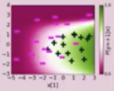
Regularize so much to avoid vanishing training loss.

# $\ell_2$ Penalty for Regularization
## Method

$\ell_2$-norm regularization or weight decay.

$$\min_{\theta \in \Theta} \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim \hat{P}_g}[\ell(\theta;(x,y))] + \lambda \|\theta\|_2$$

An example for logistic regression for classification[1].

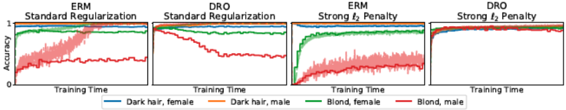
| Regularization | λ = 0 | λ = 0.00001 | λ = 0.001 | λ = 1 |
|---|---|---|---|---|
| Range of coefficients | -3170 to 3803 | -8.04 to 12.14 | -0.70 to 1.25 | -0.13 to 0.57 |
| Learned probabilities |  |  |  |  |

[1] https://www.coursera.org/lecture/ml-classification/visualizing-effect-of-l2-regularization-in-logistic-regression-1VXLD

# $\ell_2$ Penalty for Regularization

## Results

| | | | Average Accuracy | | Worst-Group Accuracy | |
|---|---|---|---|---|---|---|
| | | | ERM | DRO | ERM | DRO |
| Standard Regularization | Waterbirds | Train | 100.0 | 100.0 | 100.0 | 100.0 |
| | | Test | 97.3 | 97.4 | 60.0 | 76.9 |
| | CelebA | Train | 100.0 | 100.0 | 99.9 | 100.0 |
| | | Test | 94.8 | 94.7 | 41.1 | 41.1 |
| | MultiNLI | Train | 99.9 | 99.3 | 99.9 | 99.0 |
| | | Test | 82.5 | 82.0 | 65.7 | 66.4 |
| Strong $\ell_2$ Penalty | Waterbirds | Train | 97.6 | 99.1 | 35.7 | 97.5 |
| | | Test | 95.7 | 96.6 | 21.3 | 84.6 |
| | CelebA | Train | 95.7 | 95.0 | 40.4 | 93.4 |
| | | Test | 95.8 | 93.5 | 37.8 | 86.7 |
| Early Stopping | Waterbirds | Train | 86.2 | 80.1 | 7.1 | 74.2 |
| | | Test | 93.8 | 93.2 | 6.7 | 86.0 |
| | CelebA | Train | 91.3 | 87.5 | 14.2 | 85.1 |
| | | Test | 94.6 | 91.8 | 25.0 | 88.3 |
| | MultiNLI | Train | 91.5 | 86.1 | 78.6 | 83.3 |
| | | Test | 82.8 | 81.4 | 66.0 | 77.7 |




- ERM sacrifices worst-group training loss.
- DRO has no choice but to improve worst-group training loss.

# Early Stopping for Regularization

## Intuition (I)

**Why early stopping is a form of regularization?**

**Implicit regularization of gradient descent.**

Consider the regularized least square problem:

$$\hat{\theta}_\lambda = \arg\min_\theta \ \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

Optimal solution:

$$\hat{\theta}_\lambda = (X^\intercal X + \lambda n I)^{-1} X^\intercal Y$$

$$\underbrace{\hat{\theta}_\lambda \approx (X^\intercal X)^{-1} X^\intercal Y}_{\text{low regularization}} \qquad \underbrace{\hat{\theta}_\lambda \approx (\lambda n)^{-1} X^\intercal Y}_{\text{high regularization}}$$

Solve through gradient descent:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{\gamma}{n} X^\intercal (X\hat{\theta}_t - Y)$$

skipping a lot of details ...

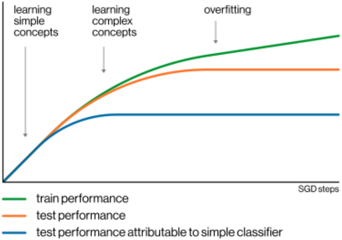
$$\underbrace{\hat{\theta}_t \approx (X^\intercal X)^{-1} X^\intercal Y}_{\text{large t}} \qquad \underbrace{\hat{\theta}_t \approx \gamma(n)^{-1} X^\intercal Y}_{\text{small t}}$$

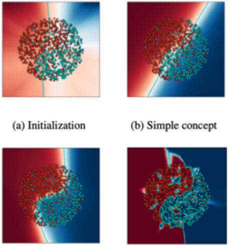
# Early Stopping for Regularization

## Intuition (II)

**SGD on Neural Networks Learns Functions of Increasing Complexity**


**Preetum Nakkiran** Harvard University
**Gal Kaplun** Harvard University
**Dimitris Kalimeris** Harvard University
**Tristan Yang** Harvard University
**Benjamin L. Edelman** Harvard University
**Fred Zhang** Harvard University
**Boaz Barak** Harvard University

(a) Initialization (b) Simple concept

(c) Complex concept (d) Overfit

# Early Stopping for Regularization

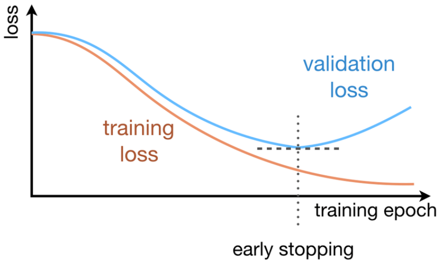
## Results

| | | | Average Accuracy | | Worst-Group Accuracy | |
|---|---|---|---|---|---|---|
| | | | ERM | DRO | ERM | DRO |
| Standard Regularization | Waterbirds | Train | 100.0 | 100.0 | 100.0 | 100.0 |
| | | Test | 97.3 | 97.4 | 60.0 | 76.9 |
| | CelebA | Train | 100.0 | 100.0 | 99.9 | 100.0 |
| | | Test | 94.8 | 94.7 | 41.1 | 41.1 |
| | MultiNLI | Train | 99.9 | 99.3 | 99.9 | 99.0 |
| | | Test | 82.5 | 82.0 | 65.7 | 66.4 |
| Strong $\ell_2$ Penalty | Waterbirds | Train | 97.6 | 99.1 | 35.7 | 97.5 |
| | | Test | 95.7 | 96.6 | 21.3 | 84.6 |
| | CelebA | Train | 95.7 | 95.0 | 40.4 | 93.4 |
| | | Test | 95.8 | 93.5 | 37.8 | 86.7 |
| Early Stopping | Waterbirds | Train | 86.2 | 80.1 | 7.1 | 74.2 |
| | | Test | 93.8 | 93.2 | 6.7 | 86.0 |
| | CelebA | Train | 91.3 | 87.5 | 14.2 | 85.1 |
| | | Test | 94.6 | 91.8 | 25.0 | 88.3 |
| | MultiNLI | Train | 91.5 | 86.1 | 78.6 | 83.3 |
| | | Test | 82.8 | 81.4 | 66.0 | 77.7 |

# Early Stopping for Regularization

## Method

**How do we usually decide to stop?** (Ignore the double descent)




Conventional approach:

- Validation sets are constructed by randomly dividing the data.

This paper:

- Validation set has balanced groups.
- Robust validation accuracy is used.

# Group adjustments

## General Idea

**Are we done? Can't we do better?**

Problem:

- The generalization gap in some groups is larger.

Solution:

- "Focus" more on groups with larger gap during training.
- Define generalization gap for each group.

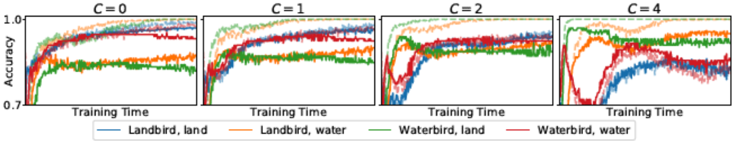
$$\delta_g = \mathbb{E}_{(x,y)\sim P_g}[\ell(\theta;(x,y))] - \mathbb{E}_{(x,y)\sim \hat{P}_g}[\ell(\theta;(x,y))]$$

$$\hat{\theta}_{\text{DRO-Adj}} := \arg\min_{\theta\in\Theta} \max_{g\in\mathcal{G}} \left\{\mathbb{E}_{(x,y)\sim \hat{P}_g}[\ell(\theta;(x,y))] + \frac{\overbrace{C}^{\text{model capacity constant}}}{\sqrt{\underbrace{n_g}_{\text{group size}}}}\right\}$$

# Group adjustments

## Results

| | Average Accuracy | | Worst-Group Accuracy | |
|---|---|---|---|---|
| | Naïve | Adjusted | Naïve | Adjusted |
| Waterbirds | 96.6 | 93.7 | 84.6 | 90.5 |
| CelebA | 93.5 | 93.4 | 86.7 | 87.8 |



# Group adjustments

## Discussion

**What might be missing in this problem formulation?**

$$\hat{\theta}_{\textsf{DRO-Adj}} := \arg\min_{\theta\in\Theta} \max_{g\in\mathcal{G}} \{\mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta;(x,y))] + \frac{\overbrace{C}^{\text{model capacity constant}}}{\sqrt{\underbrace{n_g}_{\text{group size}}}}\}$$

**They ignore the possibility that one group might be harder to generalize regardless of group size.**

$$\min_{\theta\in\Theta} \max_{g\in\mathcal{G}} \{\mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta;(x,y))] + \frac{\overbrace{\beta_g}^{\text{group complexity}} \quad \overbrace{C}^{\text{model capacity constant}}}{\sqrt{\underbrace{n_g}_{\text{group size}}}}\}$$

# Importance Weighting

Used when train and test distributions differ.

Optimize for a test distribution with uniform group frequencies.

$$\hat{\theta}_w := \arg\min_{\theta \in \Theta} \ \mathbb{E}_{(x,y,g)\sim \hat{P}_g}[w_g \ell(\theta; (x,y))]]$$

where $w_g = 1/\mathbb{E}_{g' \sim \hat{P}}[\mathbb{I}(g' = g)]$.

This paper achieves this implicitly by sampling from each group with equal probability.

| | Average Accuracy | | | Worst-Group Accuracy | | |
|---|---|---|---|---|---|---|
| | ERM | UW | DRO | ERM | UW | DRO |
| Waterbids | **97.0 (0.2)** | 95.1 (0.3) | 93.5 (0.3) | 63.7 (1.9) | 88.0 (1.3) | **91.4 (1.1)** |
| CelebA | **94.9 (0.2)** | 92.9 (0.2) | 92.9 (0.2) | 47.8 (3.7) | 83.3 (2.8) | **88.9 (2.3)** |
| MultiNLI | **82.8 (0.1)** | 81.2 (0.1) | 81.4 (0.1) | 66.4 (1.6) | 64.8 (1.6) | **77.7 (1.4)** |

# Group DRO vs. Importance Weighting

## Proposition 1

$$\boxed{\textbf{Group DRO} \stackrel{?}{=} \textbf{Importance Weighting}}$$

$$\min_{\theta \in \Theta} \mathbb{E}_{z \sim P}[w(z)\ell(\theta; z)] \quad \text{is equivalent to} \quad \min_{\theta \in \Theta} \mathbb{E}_{z \sim Q}[\ell(\theta; z)]$$

where $Q(z) \propto w(z)P(z)$.

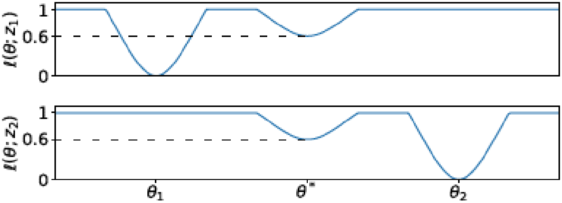
**Proposition 1.** *Suppose that the loss $\ell(\cdot; z)$ is continuous and convex for all $z$ in $\mathcal{Z}$, and let the uncertainty set $\mathcal{Q}$ be a set of distributions supported on $\mathcal{Z}$. Assume that $\mathcal{Q}$ and the model family $\Theta \subseteq \mathbb{R}^d$ are convex and compact, and let $\theta^* \in \Theta$ be a minimizer of the worst-group objective $\mathcal{R}(\theta)$. Then there exists a distribution $Q^* \in \mathcal{Q}$ such that $\theta^* \in \arg\min_\theta \mathbb{E}_{z \sim Q^*}[\ell(\theta; z)]$.*

However, this equivalence breaks down when the loss $\ell$ is non-convex:

# Group DRO vs. Importance Weighting

## Counterexample 1

**Counterexample 1.** *Consider a uniform data distribution $P$ supported on two points $\mathcal{Z} = \{z_1, z_2\}$, and let $\ell(\theta; z)$ be as in Figure 4, with $\Theta = [0, 1]$. The DRO solution $\theta^*$ achieves a worst-case loss of $\mathcal{R}(\theta^*) = 0.6$. Now consider any weights $(w_1, w_2) \in \Delta_2$ and w.l.o.g. let $w_1 \geq w_2$. The minimizer of the weighted loss $w_1\ell(\theta; z_1) + w_2\ell(\theta; z_2)$ is $\theta_1$, which only attains a worst-case loss of $\mathcal{R}(\theta^*) = 1.0$.*



# Group DRO vs. Importance Weighting

## Proof sketch

**Proposition 1.** *Suppose that the loss $\ell(\cdot; z)$ is continuous and convex for all $z$ in $\mathcal{Z}$, and let the uncertainty set $\mathcal{Q}$ be a set of distributions supported on $\mathcal{Z}$. Assume that $\mathcal{Q}$ and the model family $\Theta \subseteq \mathbb{R}^d$ are convex and compact, and let $\theta^* \in \Theta$ be a minimizer of the worst-group objective $\mathcal{R}(\theta)$. Then there exists a distribution $Q^* \in \mathcal{Q}$ such that $\theta^* \in \arg\min_\theta \mathbb{E}_{z \sim Q^*}[\ell(\theta; z)]$.*

(DRO) $\inf_{\theta \in \Theta} \mathcal{R}(\theta) = \inf_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{z \sim Q}[\ell(\theta; z)]$ is attained at $\theta^* \in \Theta$

$\sup_{Q \in \mathcal{Q}} \inf_{\theta \in \Theta} \mathbb{E}_{z \sim Q}[\ell(\theta; z)]$ is attained at $Q^* \in \mathcal{Q}$

$(\theta^*, Q^*)$ is a saddle point, i.e.,

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{z \sim Q}[\ell(\theta^*; z)] = \mathbb{E}_{z \sim Q^*}[\ell(\theta^*; z)] = \inf_{\theta \in \Theta} \mathbb{E}_{z \sim Q^*}[\ell(\theta; z)]$$

*To achieve a robust and reliable machine learning algorithm, where the model prediction does not depend on spurious correlations, we may need to focus more on worst-case generalization rather than average generalization.*