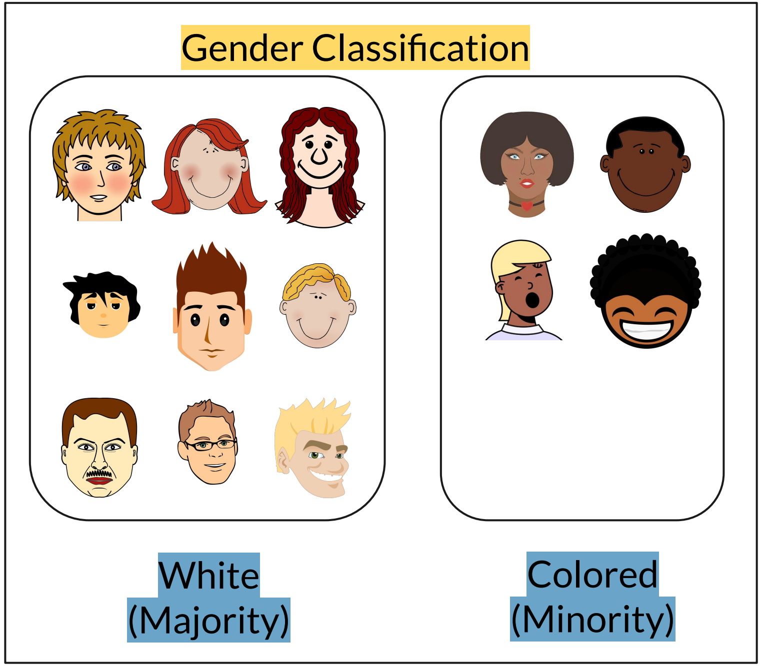
Focus on the Common Good: Group Distributional Robustness Follows

Vihari Piratla[1] Praneeth Netrapalli[2] Sunita Sarawagi[1]

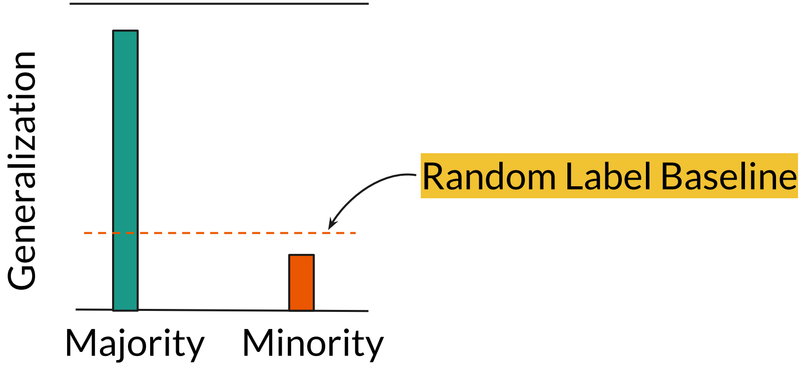
[1]IIT Bombay [2]Google Research

Datasets & Sub-population

Machine Learning datasets often contain disproportionately sized sub-populations

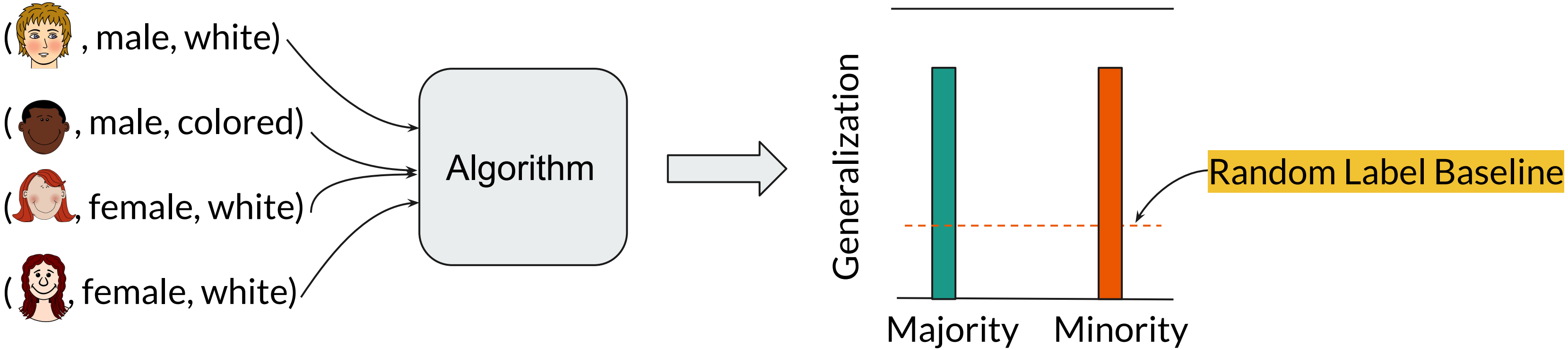


Standard methods generalize better on majority often at the expense of performance on minority.



Sub-population shift problem

Train on group annotated data with the objective of uniform generalization across all groups irrespective of their size.



Problem setting popularized by Sagawa et.al. 2020

Baselines: ERM & ERM-UW

ERM learns features from majority and overfits to minority.
Does not generalize to minority sub-population: our starting observation.

Weighted ERM (ERM-UW) up-weighs minority sub-population, improves minority's strength
Also overfits on the minority sub-population with deep models.

Baselines: Group-DRO

Group-DRO[1] trains on the group with the worst risk at any training step

- Avoids minority group overfitting since it avoids zero training loss on any group while the average loss is non-zero
- Fails when the groups have heterogeneous levels of noise or transfer as we will show

[1] Sagawa, Shiori, et al. "Distributionally robust neural networks." ICLR. 2019.

Common Gradient Descent (CGD)

💡 *Train on the group whose gradient leads to largest decrease in training loss over all groups – "common good".*

Δ_{ij} ≔ Loss decrement on j^{th} group using i^{th} group gradient
$\propto$ gradient inner product of j^{th} and i^{th} group

Goodness of i^{th} group ≔ product over all j Δ_{ij}

Pick the group with best goodness value and update parameters

CGD: algorithm

Algorithm 1 CGD Algorithm

1: **Input:** Number of groups: k, Step sizes: η_α, η
2: Initialize θ^0, $\alpha^0 = (\frac{1}{k}, \cdots, \frac{1}{k})$
3: **for** $t = 1, 2, \cdots,$ **do**
4: **for** $i \in \{1, \cdots, k\}$ **do**
5: $\Delta\ell_{is} \approx \exp(\eta_\alpha \nabla\ell_i(\theta^t)^\top \nabla\ell_s(\theta^t)) \quad \forall s \in [1 \ldots k]$
6: $\alpha_i^{t+1} \leftarrow \alpha_i^t \prod_{s \in [k]} \Delta\ell_{is}$
7: **end for**
8: $\alpha_i^{t+1} \leftarrow \alpha_i^{t+1} / \|\alpha^{t+1}\|_1 \quad \forall i \in [1 \ldots k]$
9: $\theta^{t+1} \leftarrow \theta^t - \eta \sum_{i \in \{1, \cdots, k\}} \alpha_i^{t+1} \nabla\ell_i(\theta^t)$
10: **end for**

Pick hyperparameters: η_α
α^t is t^{th} step training weight vector

For every update step and group

First order approximation of loss decrement on s^{th} group

Assimilate goodness of i^{th} group

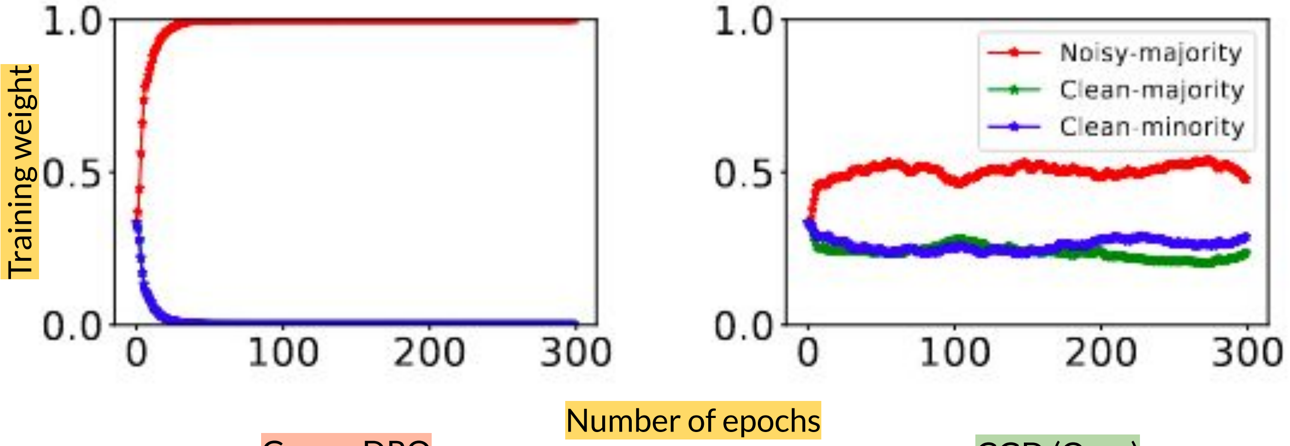
Normalize group weights, update parameters and return

🌱
We prove that CGD is a sound optimization algorithm as it converges to FOSP of macro-averaged group loss.

Synthetic Setup: Noisy Group

Randomly partitioned linearly separable data in to two majority and one minority.
First group has label noise on 20% examples.
Fit a linear classifier using Group-DRO vs CGD.

🌱
CGD is less vulnerable to noisy groups because their gradient does not transfer well.



Group-DRO

CGD (Ours)

Experiments & Evaluation

- Training data contains highly disproportionate sizes of sub-population
- Metrics: Worst and (micro) average performance on train domains
- An ideal algorithm improves worst-generalization performance without hurting average performance

Example: Blond/Non-blond on CelebA
Male-blond is 52 times smaller than male non-blond

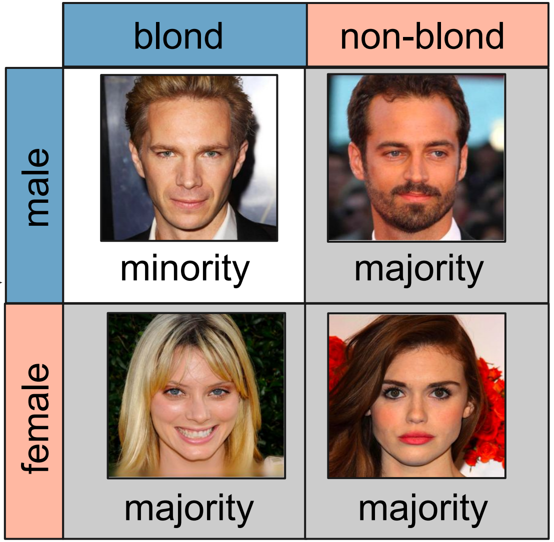


Image source: CelebA-HQ Github

Results on Standard Datasets

Standard datasets with known spurious correlations.

🌱
CGD performs as well or better than other sub-population shift, domain generalization baselines without hurting average accuracy

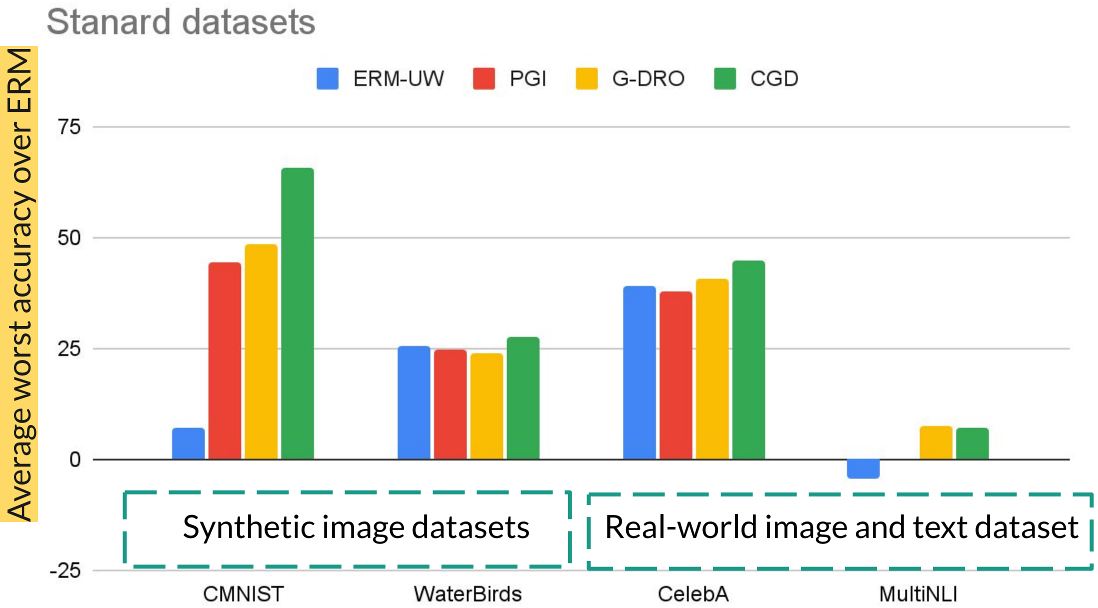
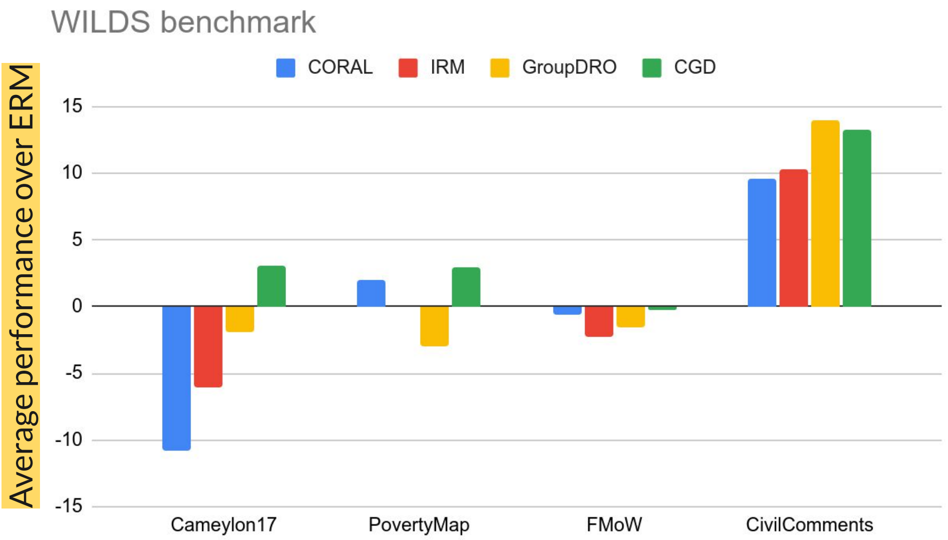


Image datasets with induced spurious correlation

PGI: Ahmed, Faruk, et al. "Systematic generalisation with group invariant predictions." *ICLR*. 2021.

CGD: Results on WILDS

1. CORAL and IRM are strong domain generalization baselines
2. Group-DRO worse than ERM on ¾ cases.
3. 🌱 CGD is at least as good as ERM and better when there is large sub-population shift.



CORAL: Sun, Baochen, and Kate Saenko. "Deep coral: Correlation alignment for deep domain adaptation." *ECCV* 2016.
IRM: Arjovsky, Martin, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).

Take-home

1. CGD is a simple new algorithm that models inter-group interaction to improve minority group generalization
2. CGD converges to stationary point of group averaged loss
3. Insights on CGD through multiple simple synthetic settings
4. On seven real-world datasets, CGD either matches or improves over strong contemporary baselines
5. Our implementation is released publicly at: https://github.com/vihari/cgd/

Thanks!