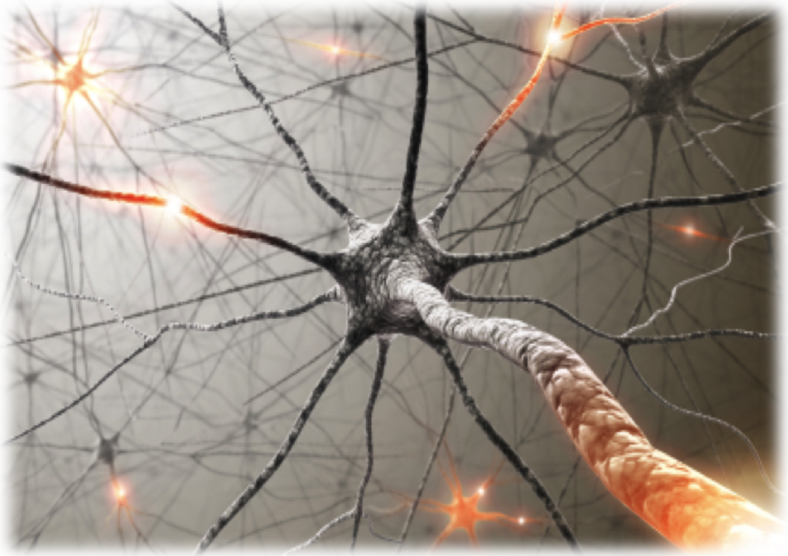
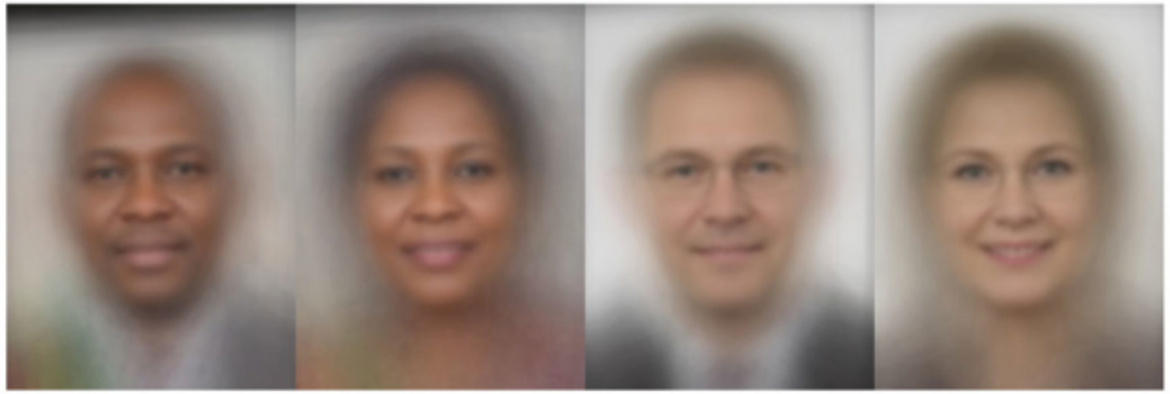
Machine learning: group DRO



[Buolamwini & Gebru 2018]

Gender Shades

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Inequalities arise in machine learning

[from a New York Times article]

False arrest due to facial recognition

Real-life consequences

Linear regression with groups

training data

	x	y	g
example	1	4	A
example	2	8	A
example	5	5	B
	6	6	B
	7	7	B
	8	8	B

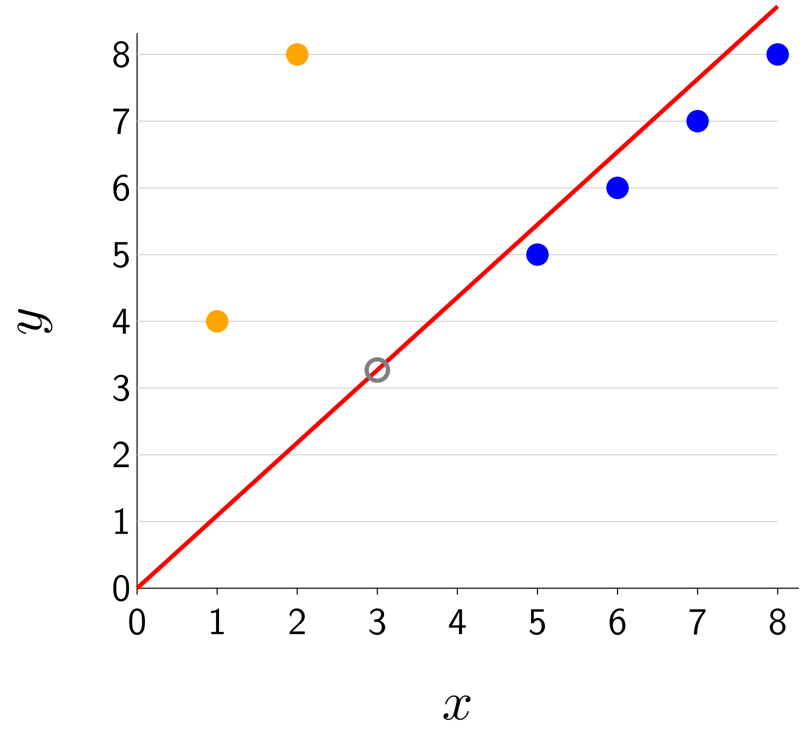
group

learning algorithm

3 input

f predictor

3.27 output



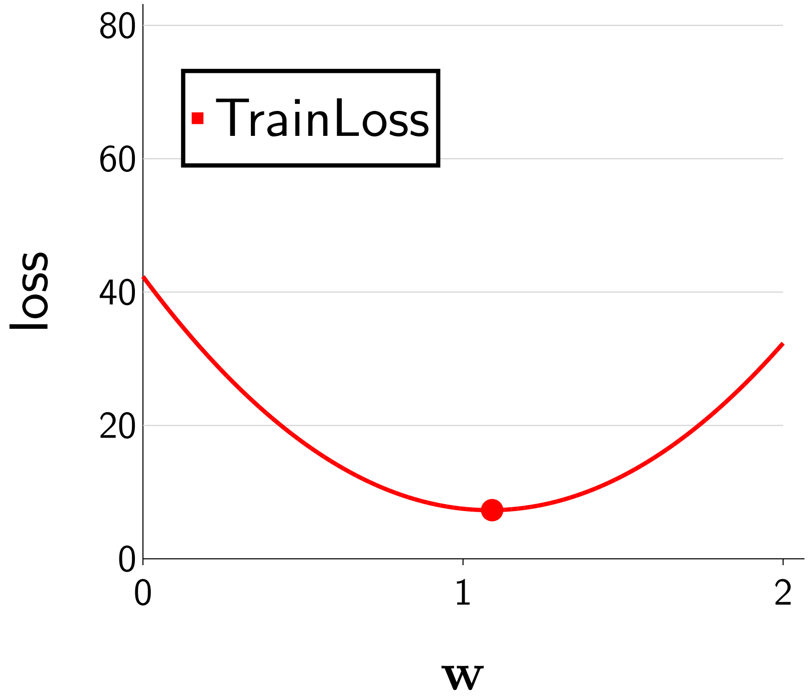
$$f_{\mathbf{w}}(x) = \mathbf{w} \cdot \phi(x) \qquad \mathbf{w} = [w] \qquad \phi(x) = [x]$$

Note: predictor $f_{\mathbf{w}}$ does not use group information g

Average loss

$$\text{Loss}(x, y, \mathbf{w}) = (f_{\mathbf{w}}(x) - y)^2$$

x	y	g
1	4	A
2	8	A
5	5	B
6	6	B
7	7	B
8	8	B



$$\text{TrainLoss}(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y) \in \mathcal{D}_{\text{train}}} \text{Loss}(x, y, \mathbf{w})$$

$\text{TrainLoss}(1) = \frac{1}{6}((1-4)^2 + (2-8)^2 + (5-5)^2 + (6-6)^2 + (7-7)^2 + (8-8)^2) = 7.5$

Per-group loss

x	y	g
1	4	A
2	8	A
5	5	B
6	6	B
7	7	B
8	8	B

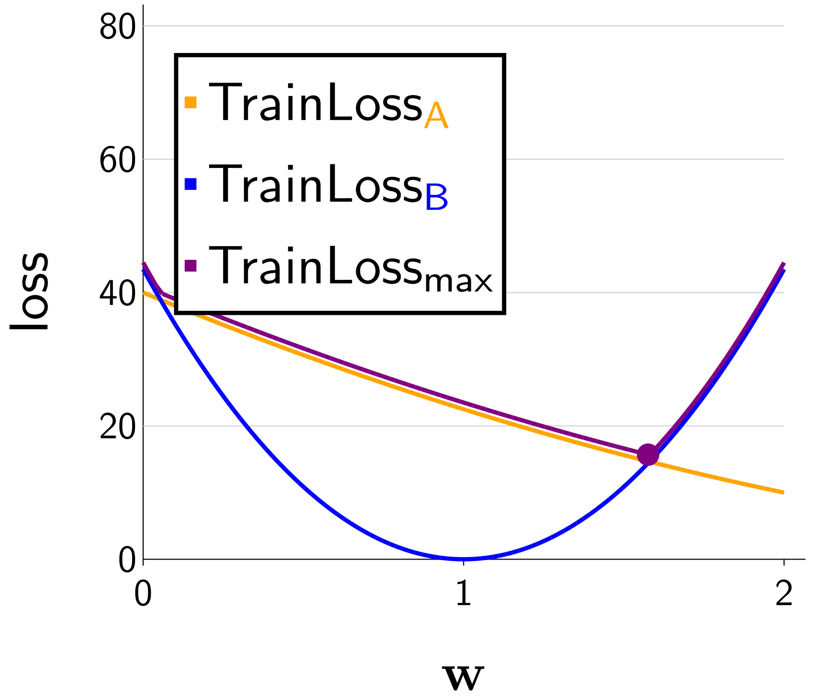
$$\text{TrainLoss}_g(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\text{train}}(g)|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}(g)} \text{Loss}(x, y, \mathbf{w})$$

$$\text{TrainLoss}_{\text{A}}(1) = \tfrac{1}{2}((1-4)^2 + (2-8)^2) = 22.5$$

$$\text{TrainLoss}_{\text{B}}(1) = \tfrac{1}{4}((5-5)^2 + (6-6)^2 + (7-7)^2 + (8-8)^2) = 0$$

Maximum group loss

x	y	g
1	4	A
2	8	A
5	5	B
6	6	B
7	7	B
8	8	B



$$\text{TrainLoss}_{\text{max}}(\mathbf{w}) = \max_g \text{TrainLoss}_g(\mathbf{w})$$

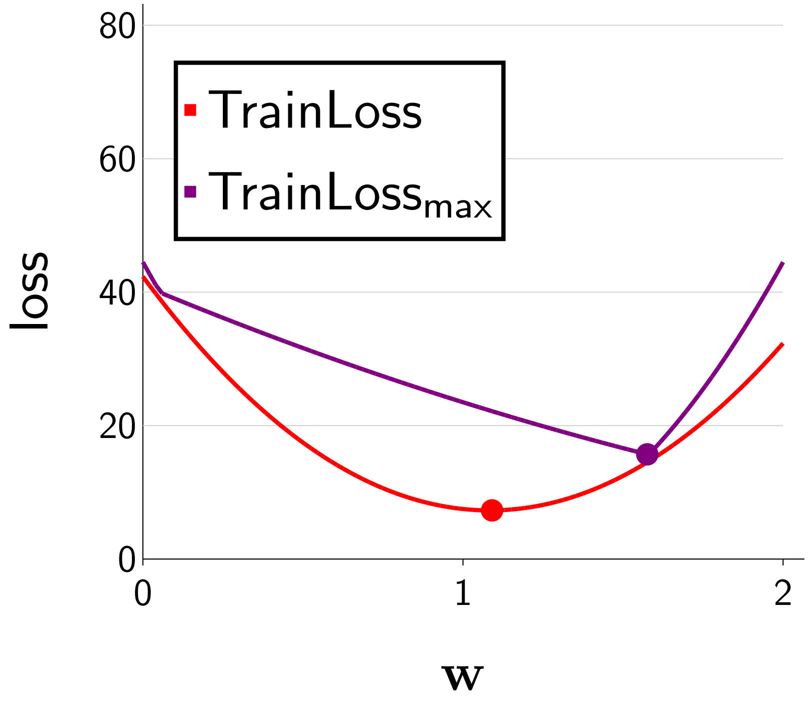
$\text{TrainLoss}_{\text{A}}(1) = 22.5$

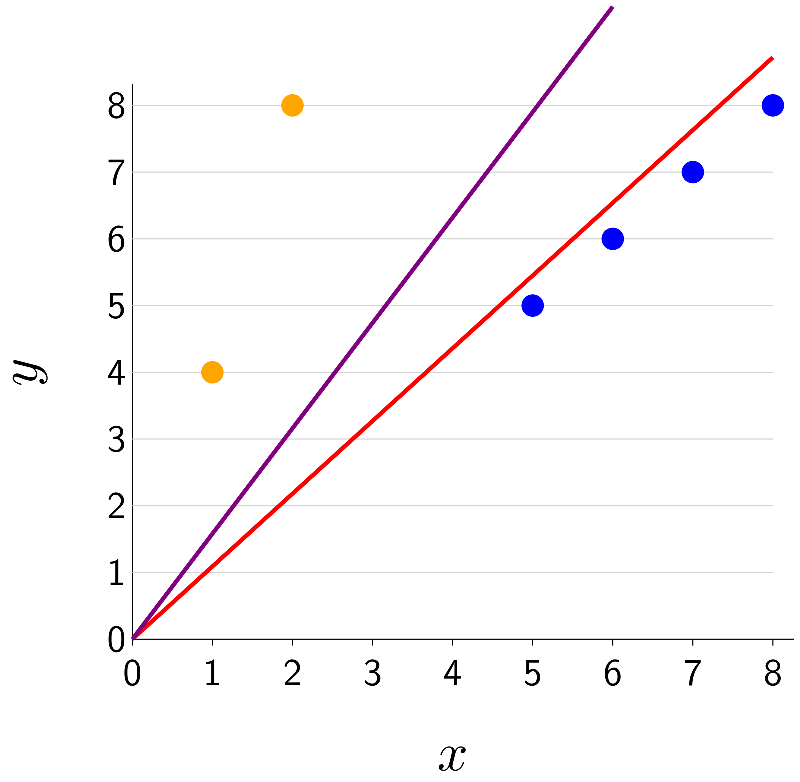
$\text{TrainLoss}_{\text{B}}(1) = 0$

$\text{TrainLoss}_{\text{max}}(1) = \max(22.5, 0) = 22.5$

Average loss versus maximum group loss

x	y	g
1	4	A
2	8	A
5	5	B
6	6	B
7	7	B
8	8	B





Standard learning:

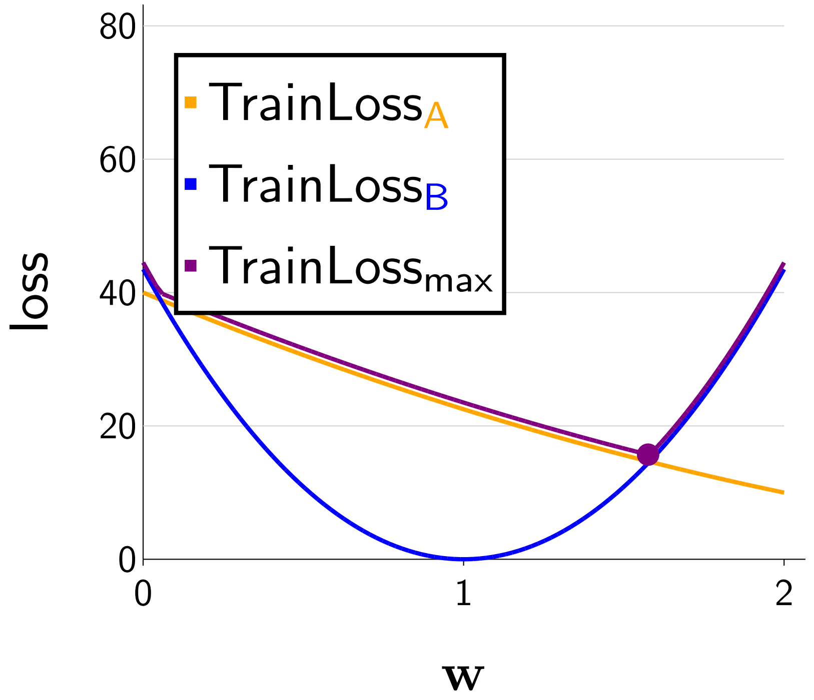
minimizer of average loss: $\mathbf{w} = 1.09$

Group distributionally robust optimization (group DRO):

minimizer of maximum group loss: $\mathbf{w} = 1.58$

Training via gradient descent

x	y	g
1	4	A
2	8	A
5	5	B
6	6	B
7	7	B
8	8	B



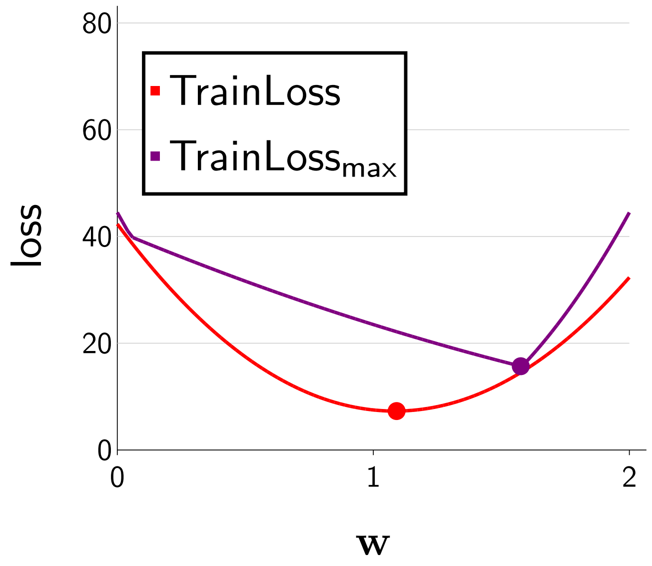
$$\text{TrainLoss}_{\text{max}}(\mathbf{w}) = \max_g \text{TrainLoss}_g(\mathbf{w})$$

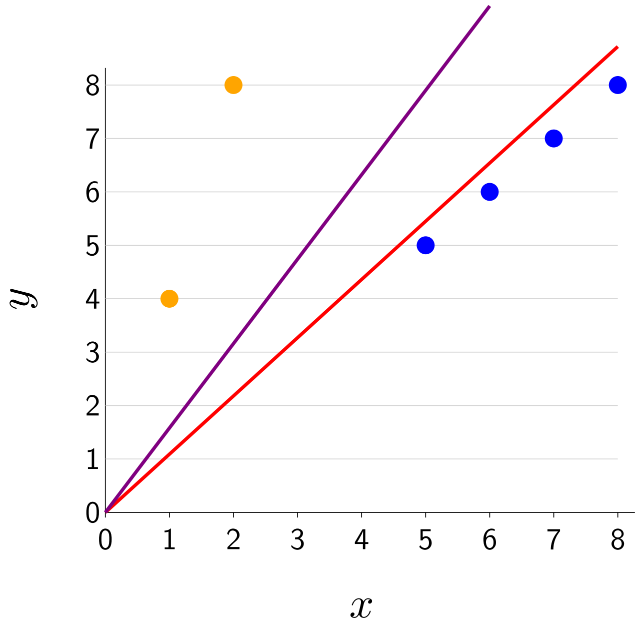
$$\nabla\text{TrainLoss}_{\text{max}}(\mathbf{w}) = \nabla\text{TrainLoss}_{g^*}(\mathbf{w})$$

$$\text{where } g^* = \arg\max_g \text{TrainLoss}_g(\mathbf{w})$$

Summary

x	y	g
1	4	A
2	8	A
5	5	B
6	6	B
7	7	B
8	8	B





- Maximum group loss $\neq$ average loss
- Group DRO: minimize the maximum group loss
- Many more nuances: intersectionality? don't know groups? overfitting?