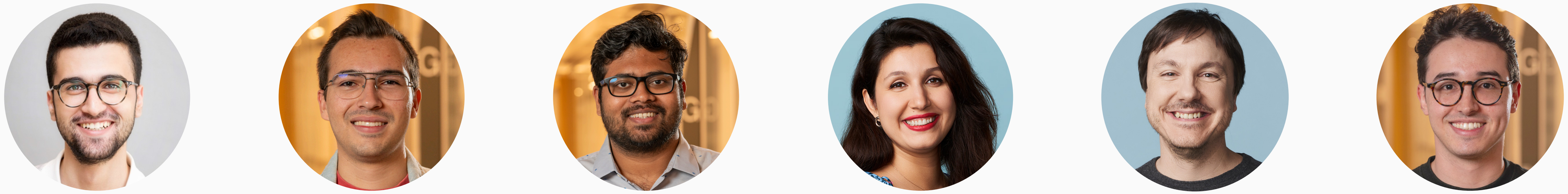
Balancing Act

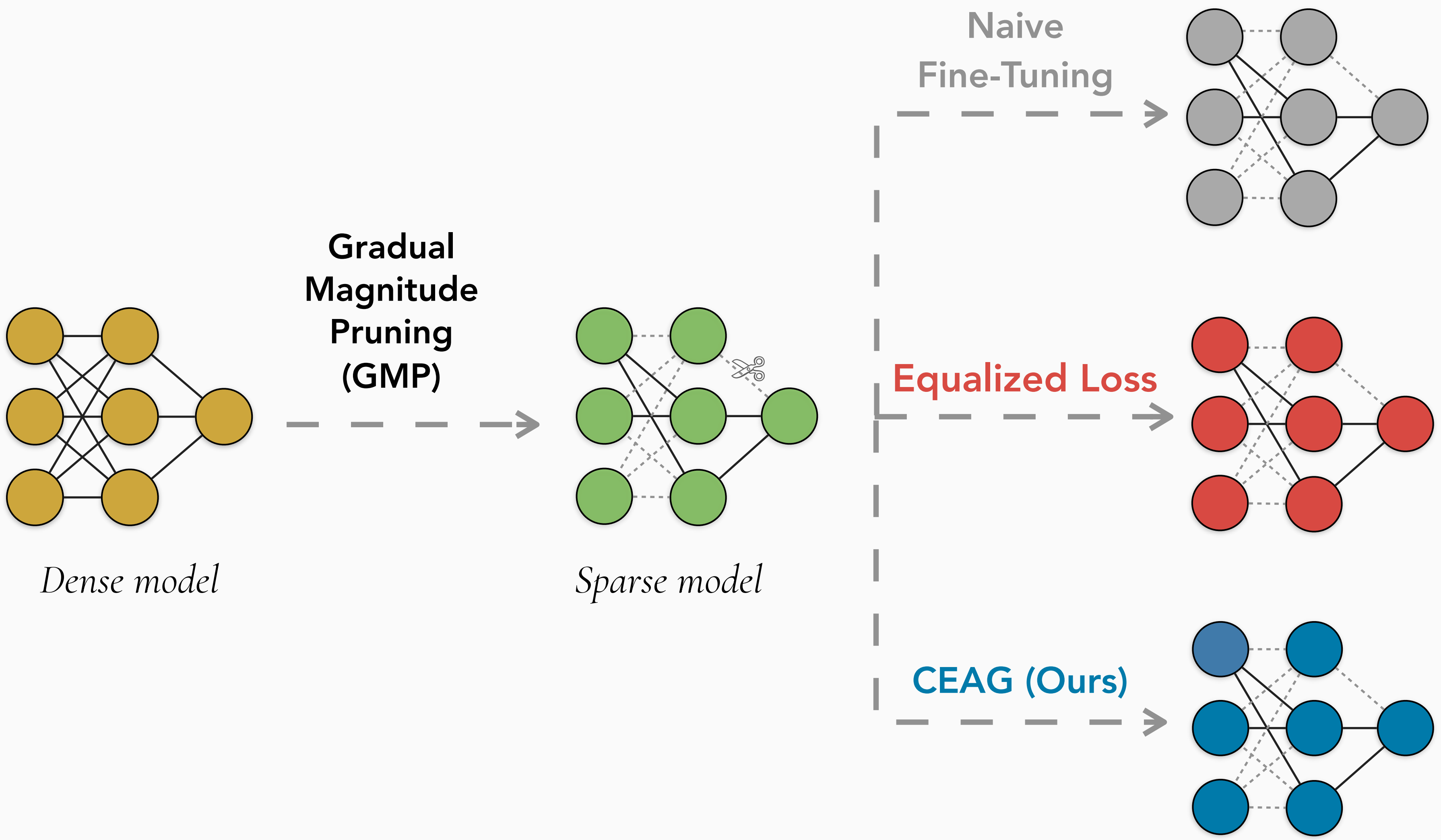
Constraining Disparate Impact in Sparse Models



Meraj Hashemizadeh, Juan Ramirez*, Rohan Sukumaran,*
Golnoosh Farnadi, Simon Lacoste-Julien and Jose Gallego-Posada

The disparate impact of pruning

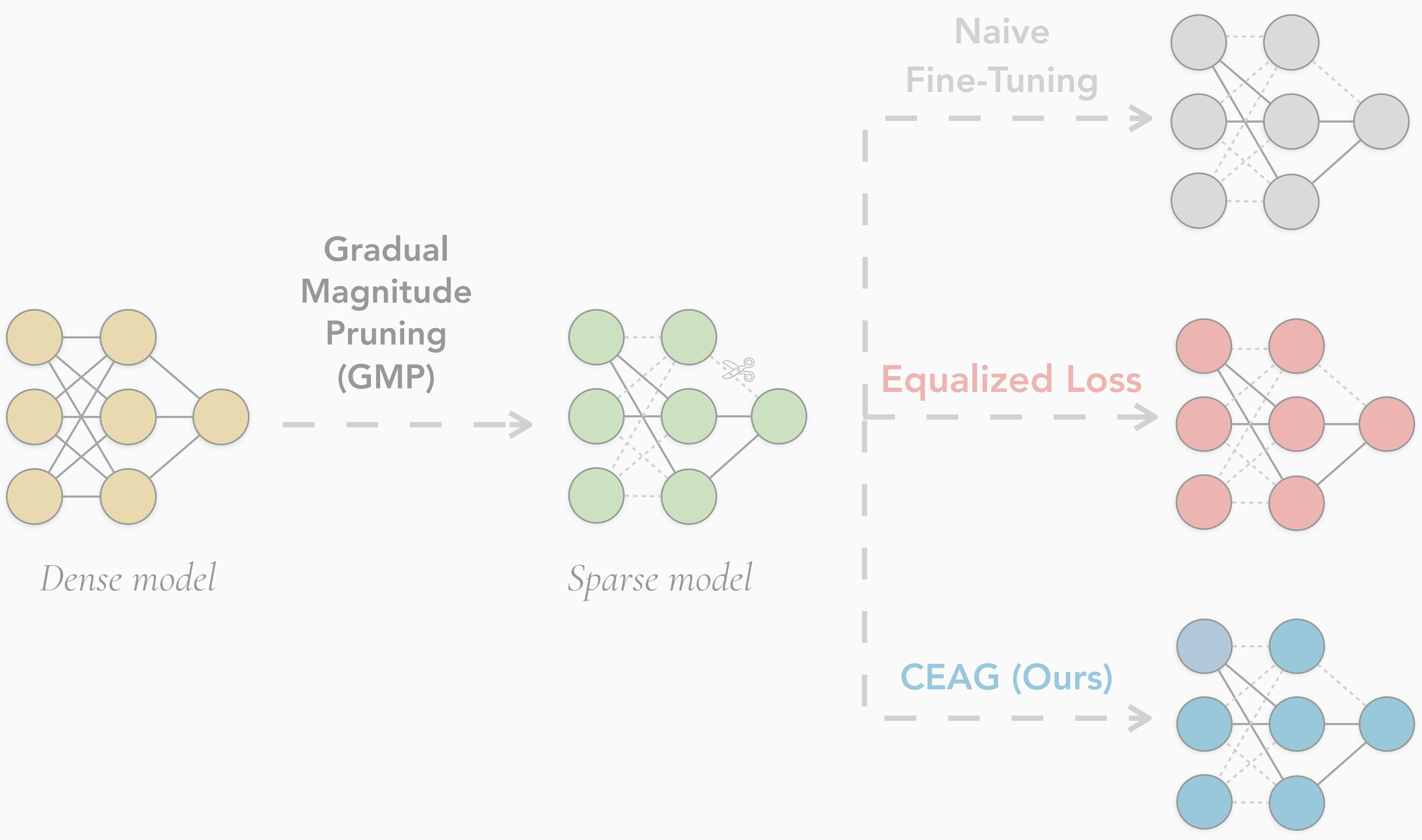
Model pruning affects the accuracy across data sub-groups unevenly

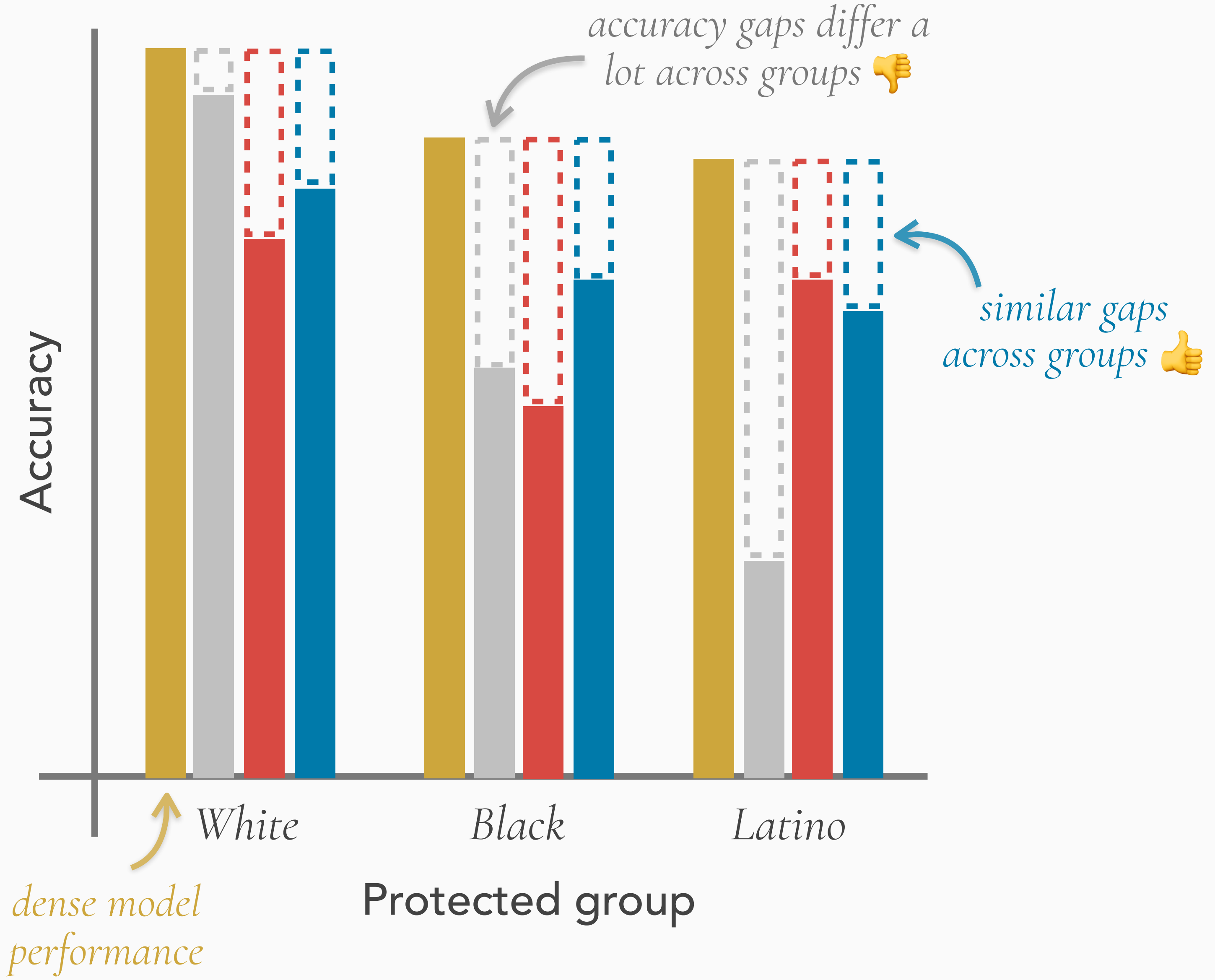


S. Hooker et al. *What Do Compressed Deep Neural Networks Forget?* 2019.
M. Paganini. *Prune Responsibly. 2020.*
C. Tran, F. Fioretto, J-E. Kim and R. Naidu. *Pruning has a disparate impact on model accuracy.* In NeurIPS, 2022.

The disparate impact of pruning

Model pruning affects the accuracy across data sub-groups unevenly





S. Hooker et al. *What Do Compressed Deep Neural Networks Forget?* 2019.
M. Paganini. *Prune Responsibly. 2020.*
C. Tran, F. Fioretto, J-E. Kim and R. Naidu. *Pruning has a disparate impact on model accuracy.* In NeurIPS, 2022.

Existing mitigation techniques

FairGRAPE (Lin et al. 2022)

Goal: minimize the variance of *accuracy drops* across groups

Approach: fairness-aware pruning, computing per-group-per-parameter importance score

Critique: scales poorly with number of groups and model size

Equalized Loss (Tran et al. 2022)

Goal: equalize *accuracy drops* across groups

Approach: equalize the per-group *losses* to the aggregate loss

Critiques:

- ignores dense model performance
- relies on the loss, a surrogate for the change in accuracy

C. Tran, F. Fioretto, J-E. Kim and R. Naidu. *Pruning has a disparate impact on model accuracy.* In NeurIPS, 2022.
X. Lin, K. Seungbae, J. Joo. *FairGRAPE: Fairness-aware GRAdient Pruning mEthod for Face Attribute Classification.* In ECCV 2022

Mitigate the disparate impact of pruning by imposing explicit constraints on the per-group accuracy changes with respect to the dense model

- 🎯 **Directly** address disparate impact by **controlling group-level accuracy changes**
- 🎯 Constraint measurements do **not rely on surrogates** (like the loss)
- 🎯 **Scale to hundreds of protected groups** and large models

Constrained Excess Accuracy Gaps problem

$$\underset{\boldsymbol{\theta}_s \in \Theta}{\text{minimize}} \; L\left(\boldsymbol{\theta}_s \,|\, \mathscr{D}_{\text{train}}\right)$$

loss of the sparse model $\boldsymbol{\theta}_s$ on the training set

$$\text{subject to} \;\; \psi_g\left(\boldsymbol{\theta}_d, \boldsymbol{\theta}_s\right) = \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}) \leq \epsilon \quad \forall g \in G$$

accuracy change on group g between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

overall accuracy change between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

Interpretability

Constraints are based on accuracy changes, and not surrogates like the loss

Flexibility

Allows for some slack in permissible deviations in accuracy gaps, given by ϵ

Accountability

Models are only acceptable if they satisfy the imposed constraints

Constrained Excess Accuracy Gaps problem

$$\underset{\boldsymbol{\theta}_s \in \Theta}{\text{minimize}} \; L\left(\boldsymbol{\theta}_s \,|\, \mathscr{D}_{\text{train}}\right)$$

loss of the sparse model $\boldsymbol{\theta}_s$ on the training set

$$\text{subject to} \;\; \psi_g\left(\boldsymbol{\theta}_d, \boldsymbol{\theta}_s\right) = \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}) \leq \epsilon \;\; \forall g \in G$$

accuracy change on group g between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

overall accuracy change between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

Interpretability

Constraints are based on accuracy changes, and not surrogates like the loss

Flexibility

Allows for some slack in permissible deviations in accuracy gaps, given by ϵ

Accountability

Models are only acceptable if they satisfy the imposed constraints

Constrained Excess Accuracy Gaps problem

$$\underset{\boldsymbol{\theta}_s \in \Theta}{\text{minimize}} \; L\left(\boldsymbol{\theta}_s \,|\, \mathscr{D}_{\text{train}}\right)$$

loss of the sparse model $\boldsymbol{\theta}_s$ on the training set

$$\text{subject to} \;\; \psi_g\left(\boldsymbol{\theta}_d, \boldsymbol{\theta}_s\right) = \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}) \leq \epsilon \;\; \forall g \in G$$

accuracy change on group g between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

overall accuracy change between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

Interpretability

Constraints are based on accuracy changes, and not surrogates like the loss

Flexibility

Allows for some slack in permissible deviations in accuracy gaps, given by ϵ

Accountability

Models are only acceptable if they satisfy the imposed constraints

Constrained Excess Accuracy Gaps problem

$$\underset{\boldsymbol{\theta}_s \in \Theta}{\text{minimize}} \; L\left(\boldsymbol{\theta}_s \,|\, \mathscr{D}_{\text{train}}\right)$$

loss of the sparse model $\boldsymbol{\theta}_s$ on the training set

$$\text{subject to} \;\; \psi_g\left(\boldsymbol{\theta}_d, \boldsymbol{\theta}_s\right) = \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}_g) - \text{Acc}(\boldsymbol{\theta}_d \,|\, \mathscr{D}) - \text{Acc}(\boldsymbol{\theta}_s \,|\, \mathscr{D}) \leq \epsilon \quad \forall g \in G$$

accuracy change on group g between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

overall accuracy change between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_s$

Interpretability

Constraints are based on accuracy changes, and not surrogates like the loss

Flexibility

Allows for some slack in permissible deviations in accuracy gaps, given by ϵ

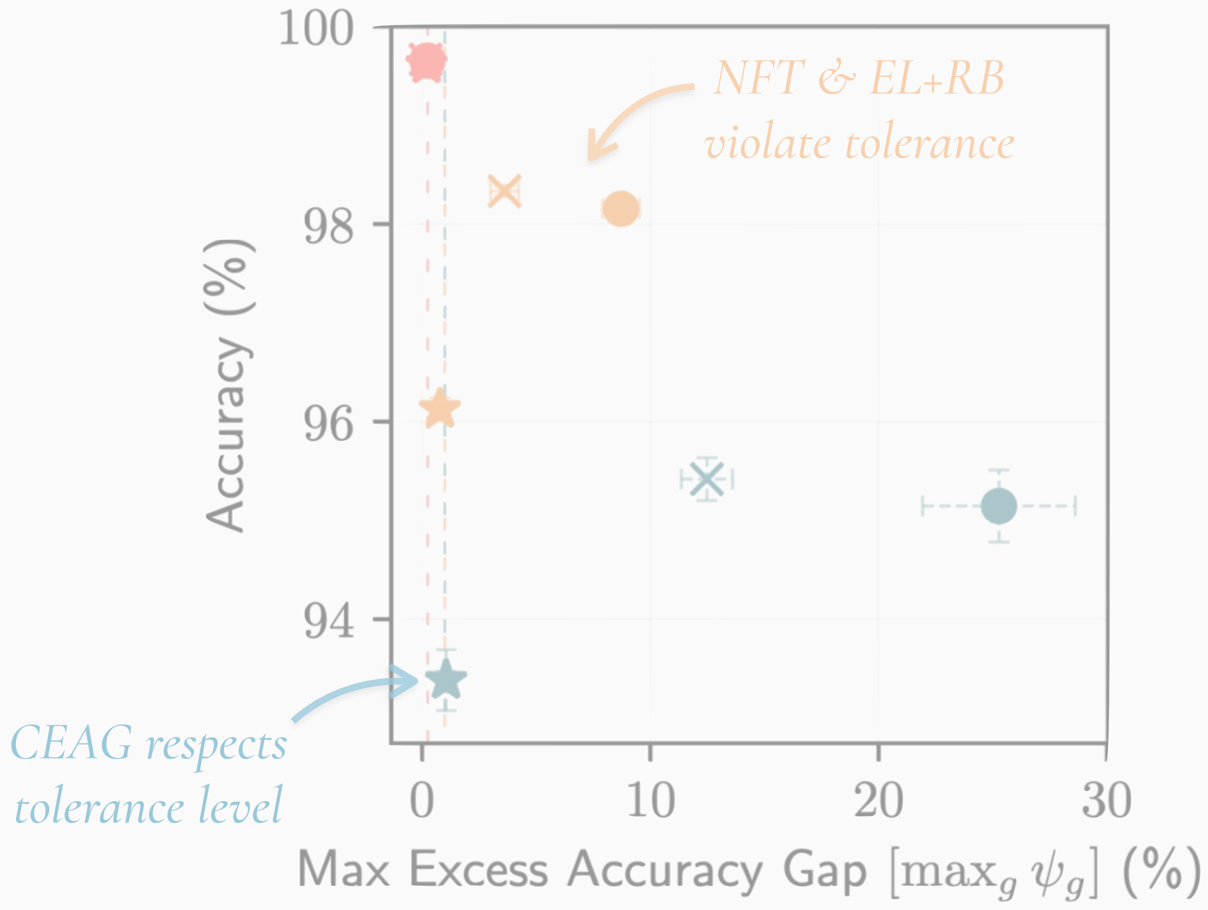
Accountability

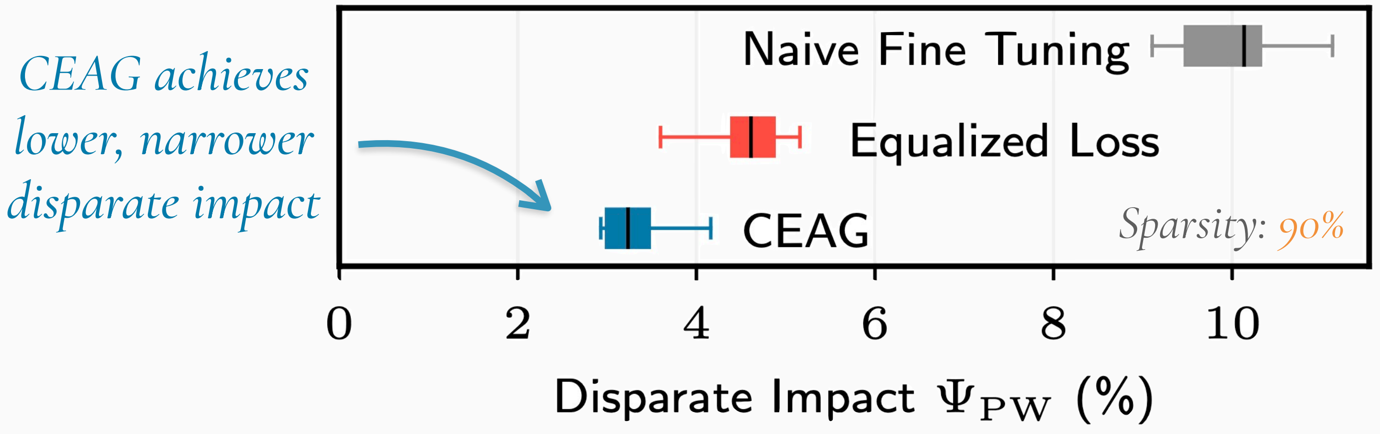
Models are only acceptable if they satisfy the imposed constraints

Results

UTKFace

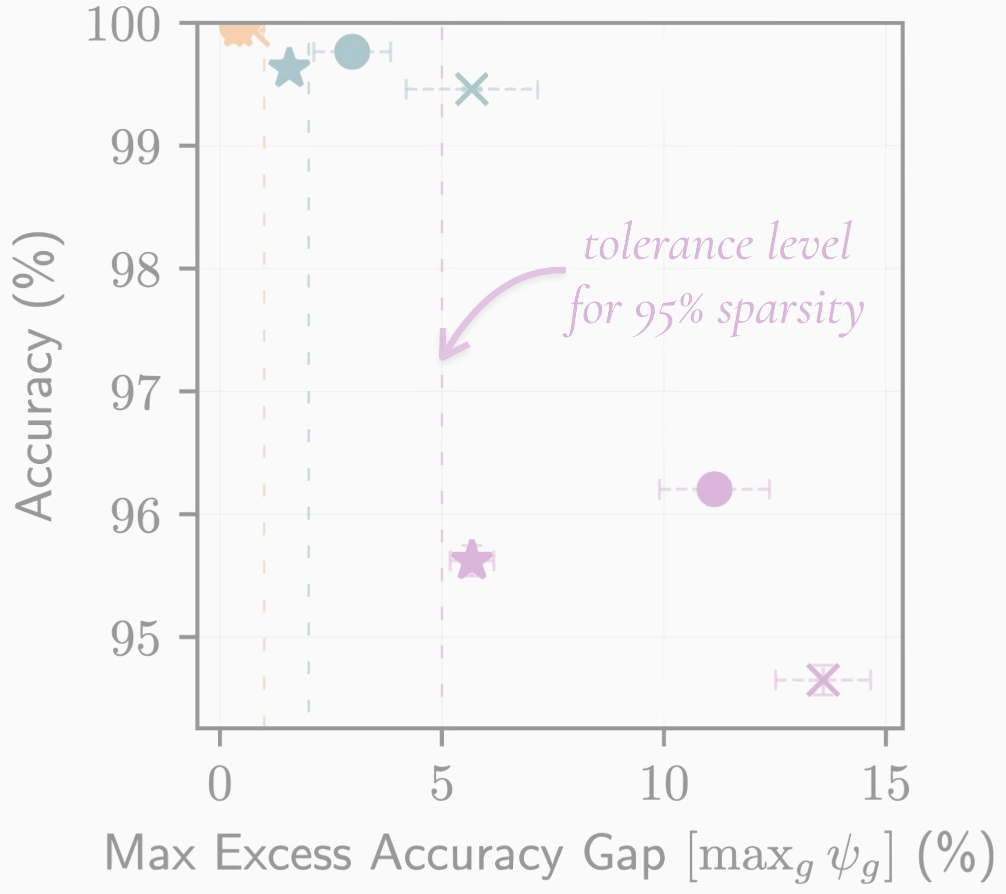
sensitive attribute: race (5 groups) — *target*: race





CIFAR100

sensitive attribute: class (100 groups) — *target*: class



⚠ *Plots show training-set metrics*

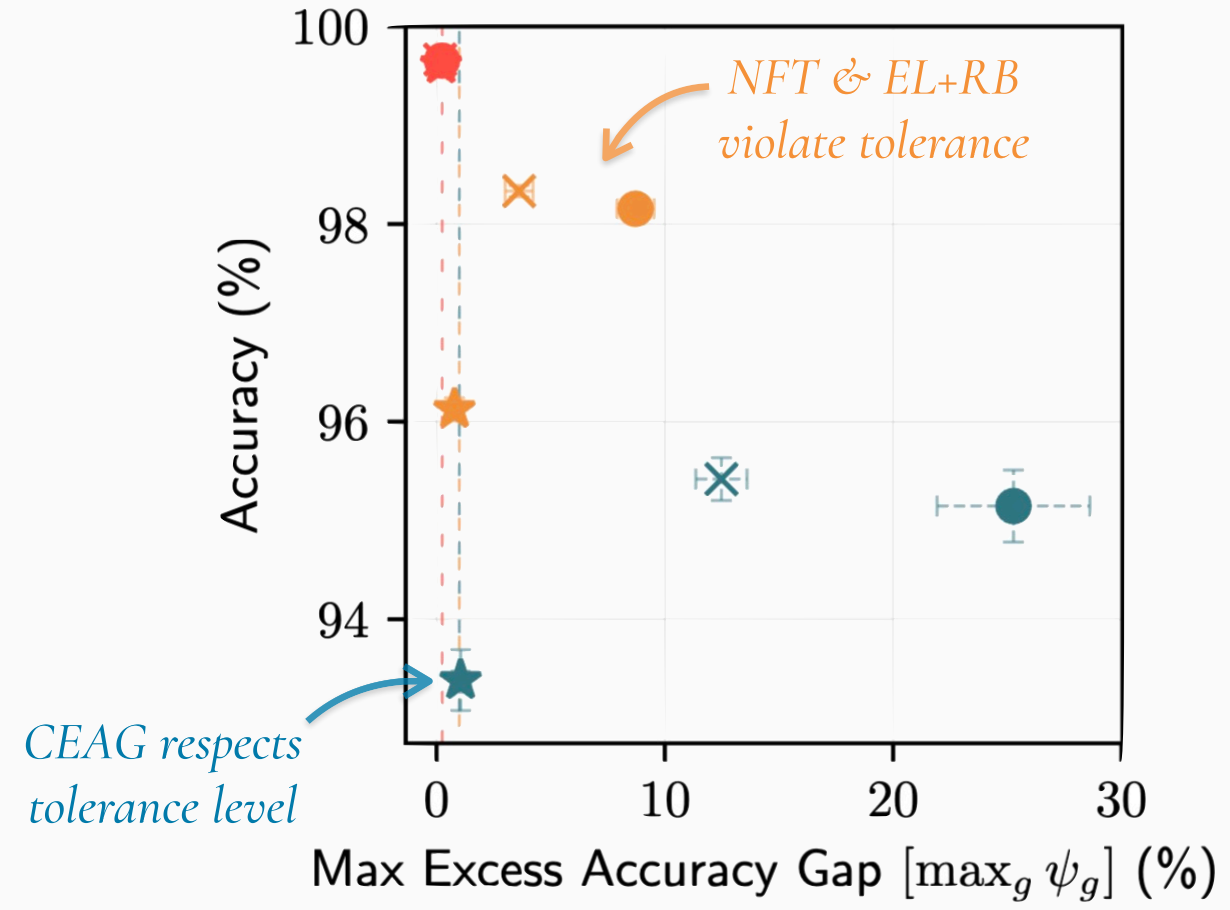
Methods: ● NFT ✕ EL + RB ★ CEAG (ours)

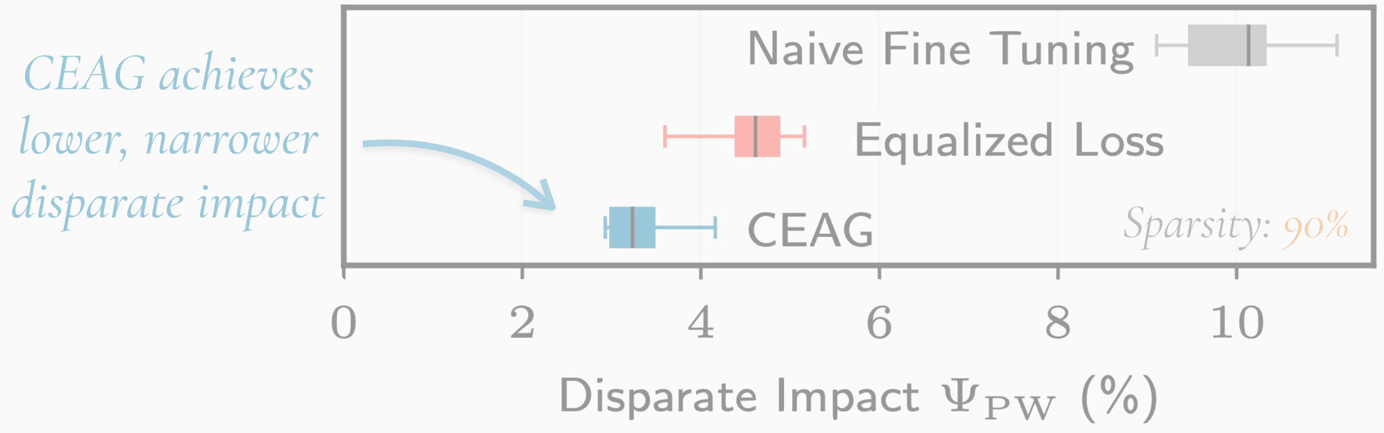
Sparsity: 85% 90% 92.5% 95%

Results

UTKFace

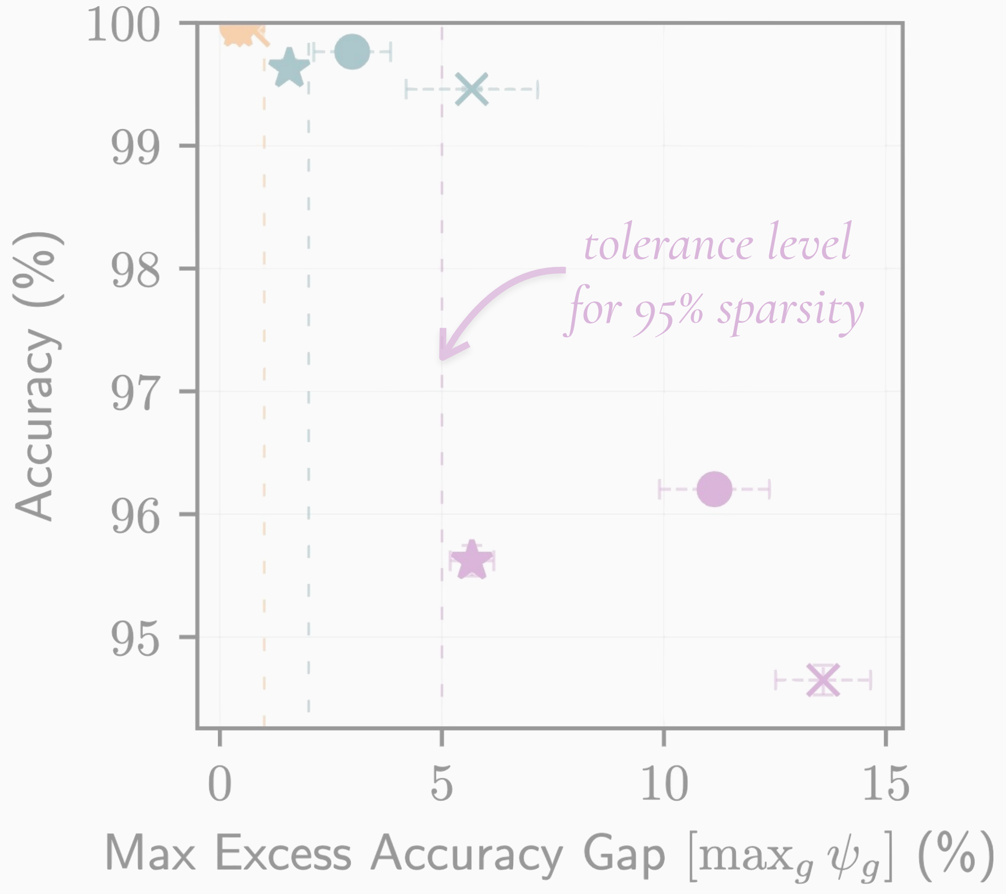
sensitive attribute: race (5 groups) — *target*: race





CIFAR100

sensitive attribute: class (100 groups) — *target*: class



⚠ *Plots show training-set metrics*

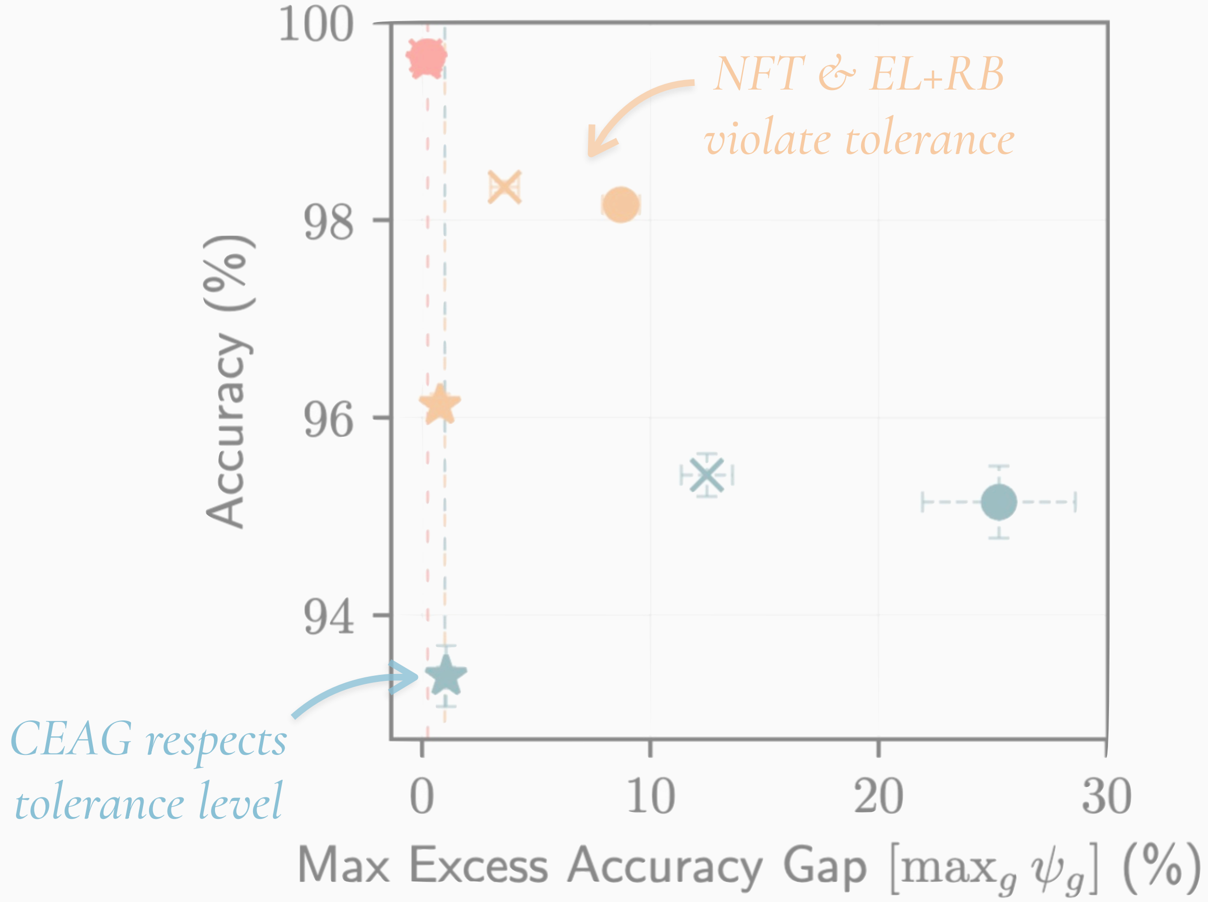
Methods: ● NFT ✕ EL + RB ★ CEAG (ours)

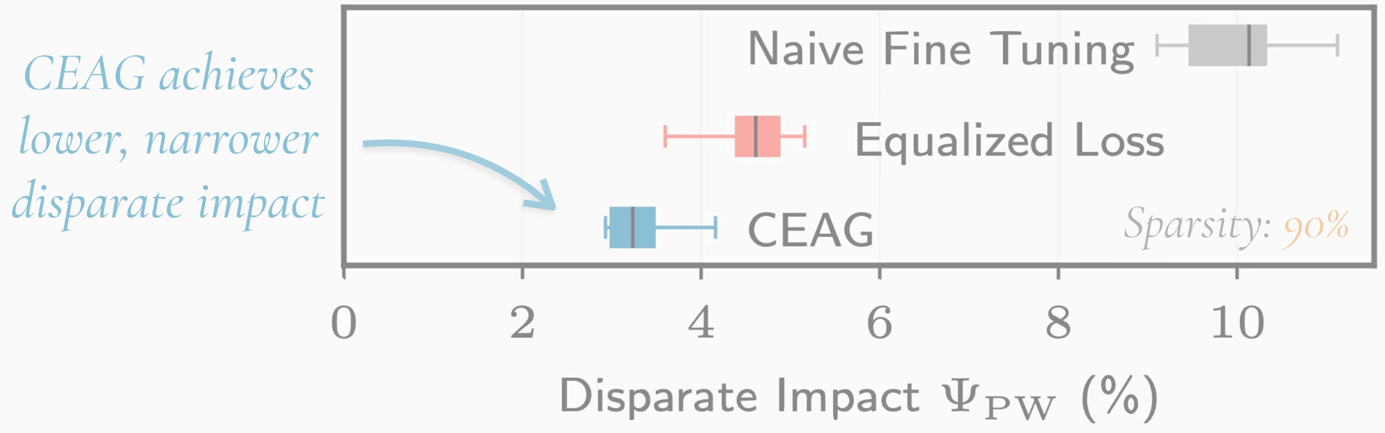
Sparsity: 85% 90% 92.5% 95%

Results

UTKFace

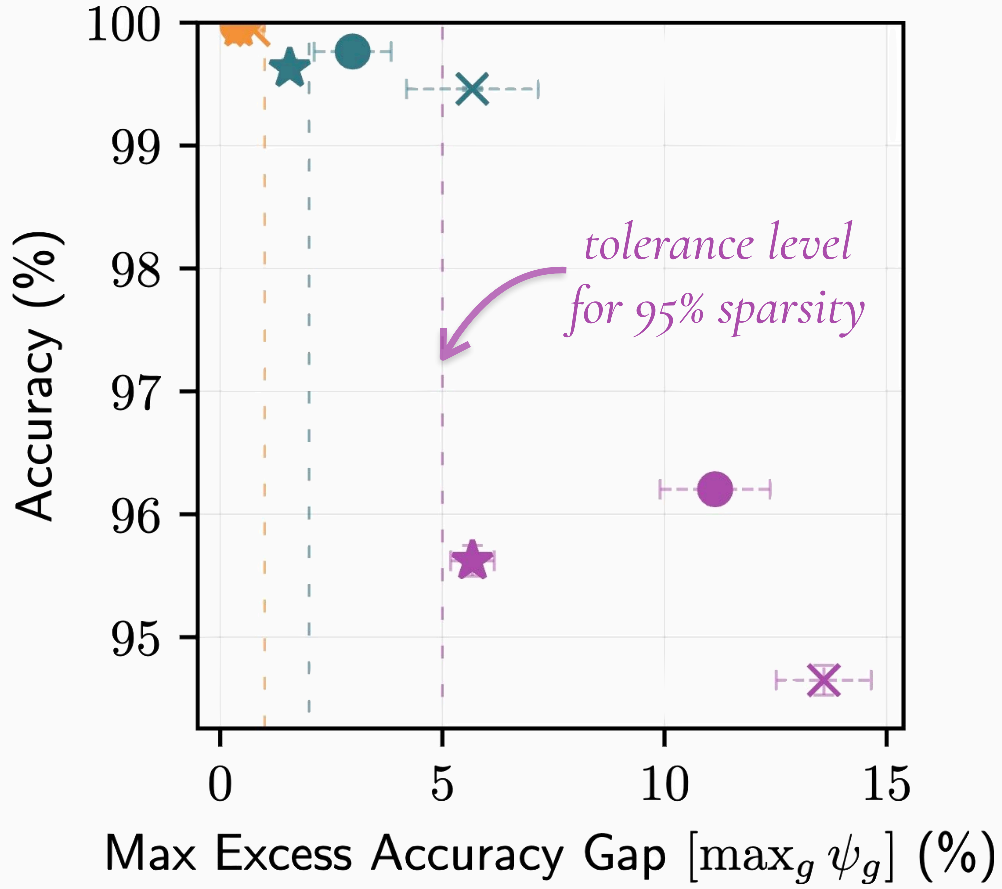
sensitive attribute: race (5 groups) — *target*: race





CIFAR100

sensitive attribute: class (100 groups) — *target*: class



⚠ *Plots show training-set metrics*

Methods: ● NFT ✕ EL + RB ★ CEAG (ours)

Sparsity: 85% 90% 92.5% 95%

The nuts & bolts

Proxy-constraints (Cotter et al. 2019)

Problem: constraints based on changes in accuracy, yielding a non-differentiable Lagrangian w.r.t. the model

Approach: use a surrogate function for computing constraint gradients, but keep non-differentiable measurement for assessing constraint satisfaction

Replay buffers (Mnih et al. 2019)

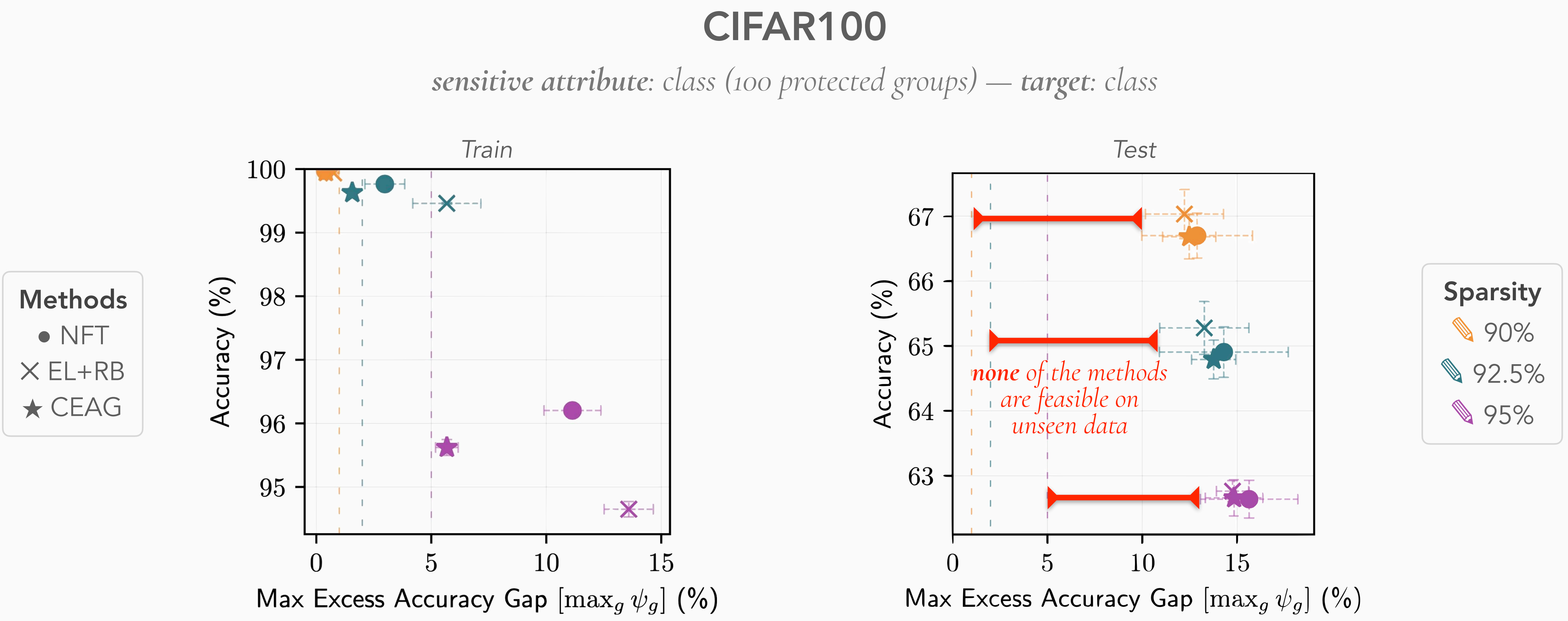
new!
can be used in any optimization problem with stochastically-estimated constraints

Problem: mini-batch estimates of the constraints can have large variance, especially for small groups

Approach: estimate the accuracy of the sparse model based on (cached predictions) on the k most recent datapoints of each group

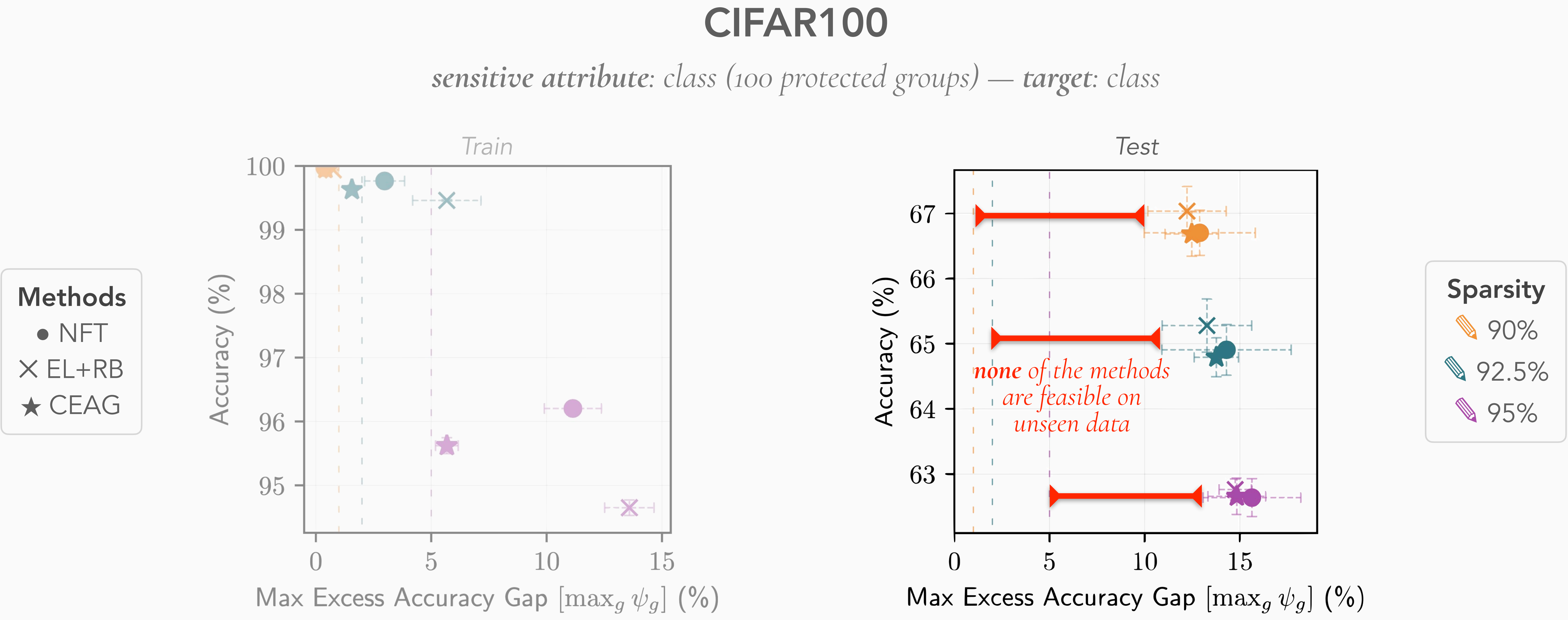
A. Cotter et al. *Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals.* In JMLR, 2019.
V. Mnih et al. *Playing Atari with Deep Reinforcement Learning.* In NeurIPS Deep Learning Workshop, 2013.

Generalization challenges



⚠️ The generalization challenge affects all surveyed methods, including ours!

Generalization challenges



⚠️ **The generalization challenge affects all surveyed methods, including ours!**

Our proposed CEAG approach

1. enables *direct mitigation* of pruning-induced disparate impact,
2. exploits *buffers* for reducing variance in constraint estimation,
3. highlights need for *further research on the test-time success* of mitigation methods,
4. showcases the use of *Cooper*—our companion library for constrained optimization in PyTorch.

Balancing Act

Constraining Disparate Impact in Sparse Models






*Meraj Hashemizadeh**

*Juan Ramirez**

Rohan Sukumaran

Golnoosh Farnadi

Simon Lacoste-Julien

Jose Gallego-Posada

ICLR

Poster session #4

Wed. May 8, 4:30 PM