The dynamics of bias in SGD training

Stefano Sarao Mannelli

Cargese 2025

CHALMERS
UNIVERSITY OF TECHNOLOGY

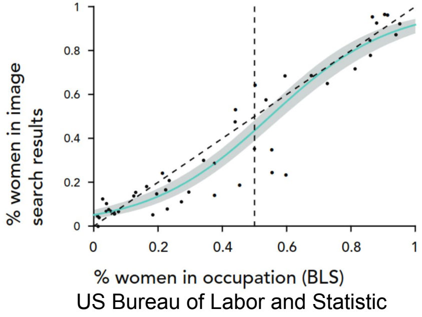
GÖTEBORG
UNIVERSITY

“data is biased” is not enough

Google images query: *“CEO”*
[Kay, et al. 2015; Megan Garcia 2017; Feng, Shah 2022]

HISTORICAL BIAS

Query	Result W%	Real W%
CEO	11%	22%
Developer	15%	20%



“data is biased” is not enough

The ML pipeline is composed of many elements that can contribute to bias generation/amplification [Suresh, Guttag 2021].

- ➢ [Sagawa, et al. 2020] : over-parameterisation increases bias,
- ➢ [SSM, et al. 2022; Jain, et al. 2024; Subramonian, et al. 2025] : data structure,
- ➢ [Bell, et al. 2023; Bell, et al. 2024] : architecture complexity
- ➢ [Iofinova, et al. 2023] : pruning can increase bias,
- ➢ [Ganesh, et al. 2023] : batch randomness and curricula,
- ➢ [Francazi, et al. 2023a;b] : architecture complexity/activation function.

Motivation

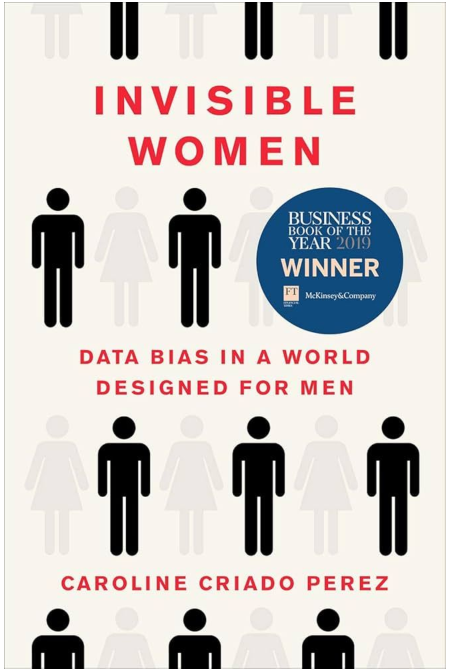
Understanding bias as a physicist

The dynamics of bias

Testing the results in the wild-ish

Conclusions

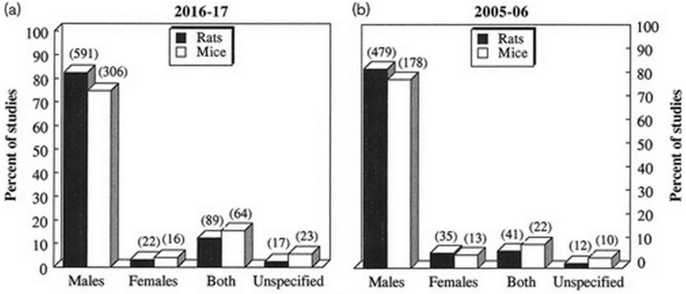
A modelling approach



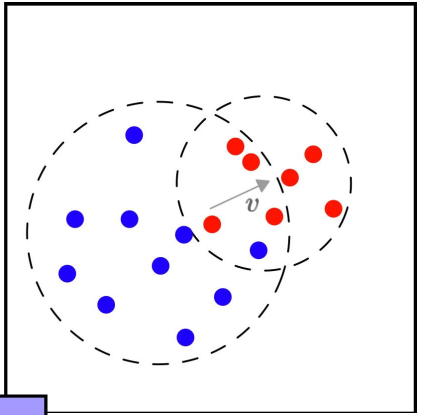
[Criado Perez 2019]

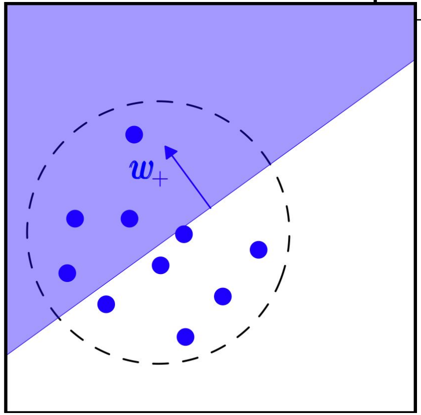
Hughes 2007; 2019 made a meta analysis of publications in pharmacology journals

- ➢ Behavioural Brain Research,
- ➢ Behavioural Pharmacology,
- ➢ Pharmacology,
- ➢ Biochemistry and Behavior,
- ➢ Physiology and Behavior,
- ➢ Psychopharmacology



A modelling approach





Gaussian Mixture model

0. Generate centroids

$$v_i \sim \mathcal{N}(0,1)$$

1. Assign group

$$c \sim \rho\delta(c-1) + (1-\rho)\delta(c+1)$$

2. Generate sample

$$x = c\frac{v}{\sqrt{N}} + z \qquad z \sim \mathcal{N}(0,1)$$

(Single Index) Teacher-Student model

0. Generate teacher

$$W_T \in \mathbb{S}^{N-1}(N)$$

1. Generate sample

$$x \sim \mathcal{N}(0,1)$$

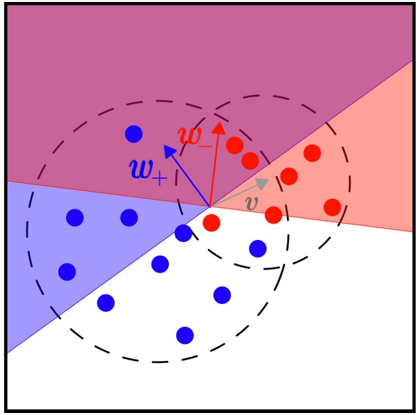
2. Generate label

$$y = \mathrm{sign}\left(\frac{x \cdot W_T^c}{\sqrt{N}} + b_c\right)$$

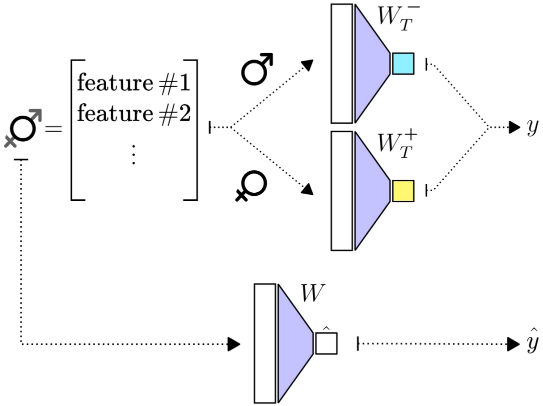
A modelling approach

> **Rmk.** In high-dimension the key observables can be characterised by a few sufficient statistics
>
> $$Q = \frac{1}{N} W \cdot W \qquad M = \frac{1}{N} v \cdot W$$
> $$q_T = \frac{1}{N} W_T^- \cdot W_T^+ \qquad R^{\pm} = \frac{1}{N} W_T^{\pm} \cdot W$$

[**SSM**, et al. 2022]



Teacher-Mixture model



0. Generate teachers and centroids

$$W_T \in \mathbb{S}^{N-1}(N) \qquad v_i \sim \mathcal{N}(0,1)$$

1. Assign group

$$c \sim \rho\delta(c-1) + (1-\rho)\delta(c+1)$$

2. Generate sample

$$x = c\frac{v}{\sqrt{N}} + z \qquad z \sim \mathcal{N}(0,1)$$

3. Generate label

$$y = \text{sign}\left(\frac{x \cdot W_T^c}{\sqrt{N}} + b_c\right)$$

4. Train

$$\mathcal{L}(\boldsymbol{W}, b) = \sum_{\mu \in \mathcal{D}} \ell\left(\boldsymbol{W}, b; \boldsymbol{x}^{\mu}, y^{\mu}\right) + \frac{\lambda ||\boldsymbol{W}||_2^2}{2}$$

5. Do Replicas

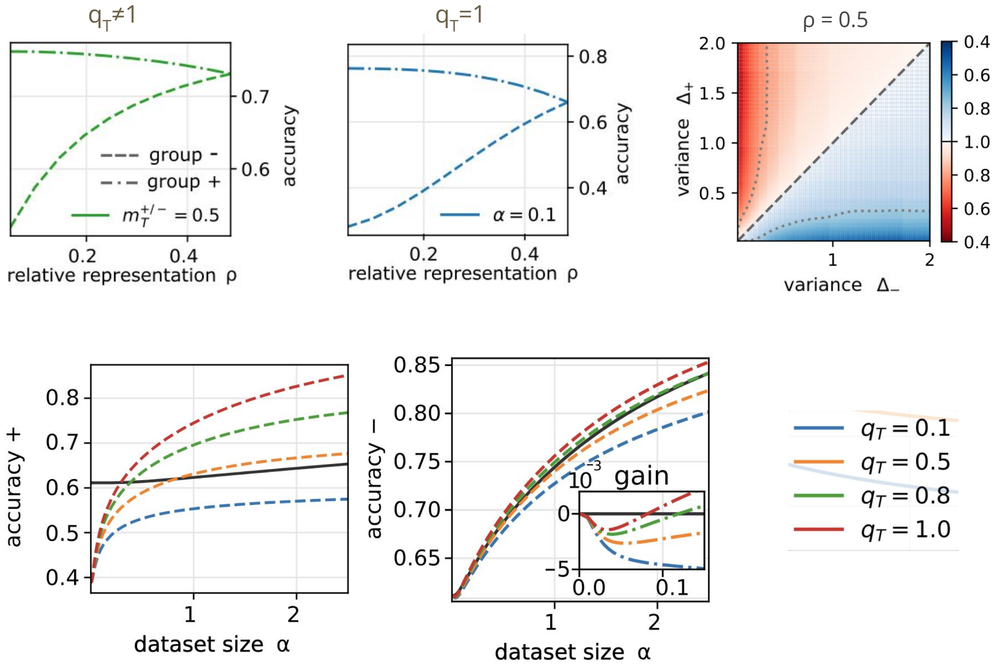
$$Q = -2\frac{\partial s(\hat{\Theta};\lambda)}{\partial\, \delta\hat{Q}}; \quad M = \frac{\partial s(\hat{\Theta};\lambda)}{\partial\, \hat{M}}; \quad R^{\pm} = \frac{\partial s(\hat{\Theta};\lambda)}{\partial\, \hat{R}^{\pm}};$$

with $s(\hat{\Theta};\lambda)$ the free entropy of the model.

Results at equilibrium (brief summary)

In [SSM, et al. 2022] we showed:

1. That bias can emerge in
   a. Model mismatched situations
   b. In model matching situations
   c. In balanced datasets
2. Despite its disadvantages, joint training is usually advantageous.



Motivation

Understanding bias as a physicist

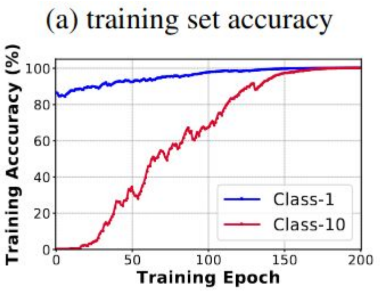
The dynamics of bias

Testing the results in the wild-ish

Conclusions

Bias evolution and mitigation strategies

A typical learning curve:

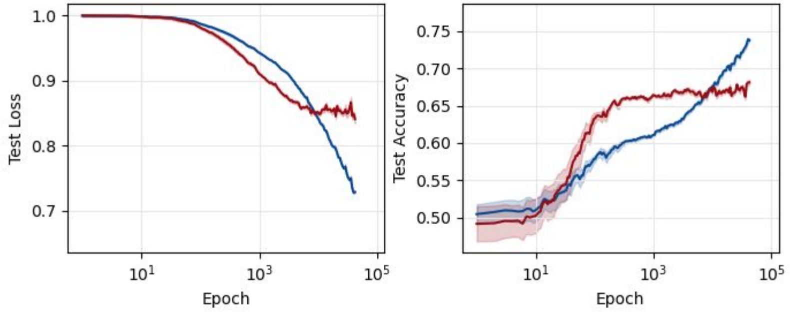


[Ye et al. 2021]

...however, is this assumption correct?

Let's revisit the experiment by Bell & Sagun 2022:

- a 2-layer NN trained online on CIFAR10 with MSE,
- divide the dataset in two populations and assign labels +1 and -1:
  - group 1: ['deer', 'bird', 'frog', 'horse']
  - group 2: ['cat', 'airplane', 'automobile', 'truck']
- Put the dataset back together with a mixture 90%-10% of group 1 and group 2.

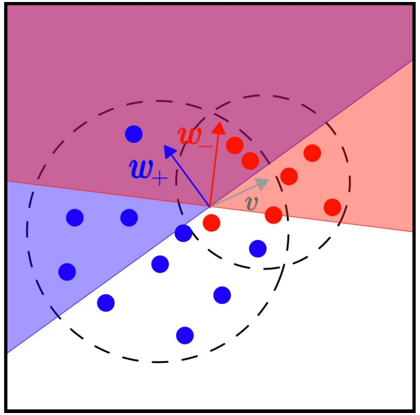


A modelling approach for oSGD

$$Q = \frac{1}{N} W \cdot W \qquad M = \frac{1}{N} v \cdot W$$

$$q_T = \frac{1}{N} W_T^- \cdot W_T^+ \qquad R^\pm = \frac{1}{N} W_T^\pm \cdot W$$

[SSM, et al. 2022]



[Saad & Solla 1995; Biehl & Schwarze 1995] MF analysis of online SGD:

$$\boldsymbol{W}[\mu+1] = W[\mu] - \frac{\eta}{\sqrt{N}} \sigma'(\lambda^\mu)(\sigma(\lambda^\mu) - y^\mu)\boldsymbol{x}^\mu$$

with $\lambda^\mu = \frac{\boldsymbol{W} \cdot \boldsymbol{x}^\mu}{\sqrt{N}}$

Extended and made rigorous by [Goldt et al. 2019] and further generalised by [Veiga et al. 2022; Arnaboldi et al. 2023; Ben Arous, Gheissari, Jagannath 2024; Collins-Woodfin, Paquette, Paquette, Seroussi 2024]...
In our case we will focus on the perceptron model.

In the high-d limit, the OPs evolve following a set of deterministic ODEs, e.g.

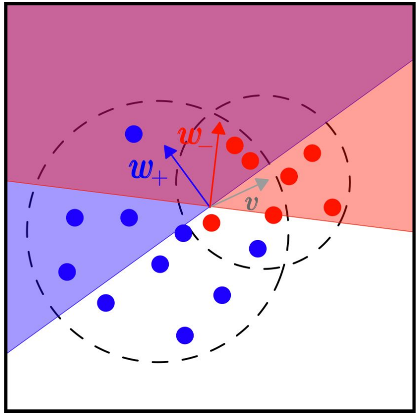
$$\frac{dR}{dt}(t) = -\eta \mathbb{E}_{\lambda,\lambda^*}\left[\sigma'(\lambda)(\sigma(\lambda) - \text{sign}(\lambda^*))\lambda^*\right] = -\eta \mathbb{E}_{\lambda,\lambda^*}\left[(\lambda - \text{sign}(\lambda^*))\lambda^*\right]$$

assume linear system

Analytical solution of the dynamics

$$Q = \frac{1}{N} W \cdot W \qquad M = \frac{1}{N} v \cdot W$$
$$q_T = \frac{1}{N} W_T^- \cdot W_T^+ \qquad R^\pm = \frac{1}{N} W_T^\pm \cdot W$$

[**SSM**, et al. 2022]



Focus on a linear classifier, the dynamics of the **order parameters** follow the ODEs below

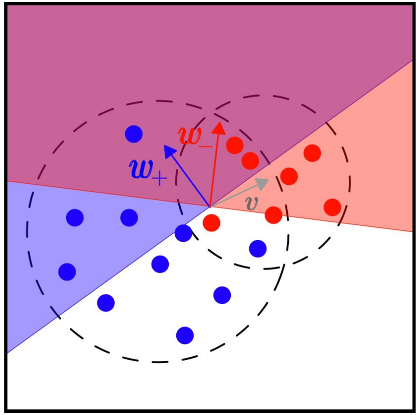
$$\frac{dQ}{dt} = c_6 + c_7 M + c_8 M^2 + c_{9+} R_+ + c_{9-} R_- + c_{10} Q$$
$$\frac{dM}{dt} = c_1 + c_2 M$$
$$\frac{dR_-}{dt} = c_{3-} + c_{4-} M + c_{5-} R_-$$
$$\frac{dR_+}{dt} = c_{3+} + c_{4+} M + c_{5+} R_+$$

with a series of coefficients **c**.

Analytical solution of the dynamics

$$Q = \frac{1}{N} W \cdot W \qquad M = \frac{1}{N} v \cdot W$$
$$q_T = \frac{1}{N} W_T^- \cdot W_T^+ \qquad R^\pm = \frac{1}{N} W_T^\pm \cdot W$$

[**SSM**, et al. 2022]



Focus on a linear classifier, the dynamics of the **order parameters** admit **explicit solution**!

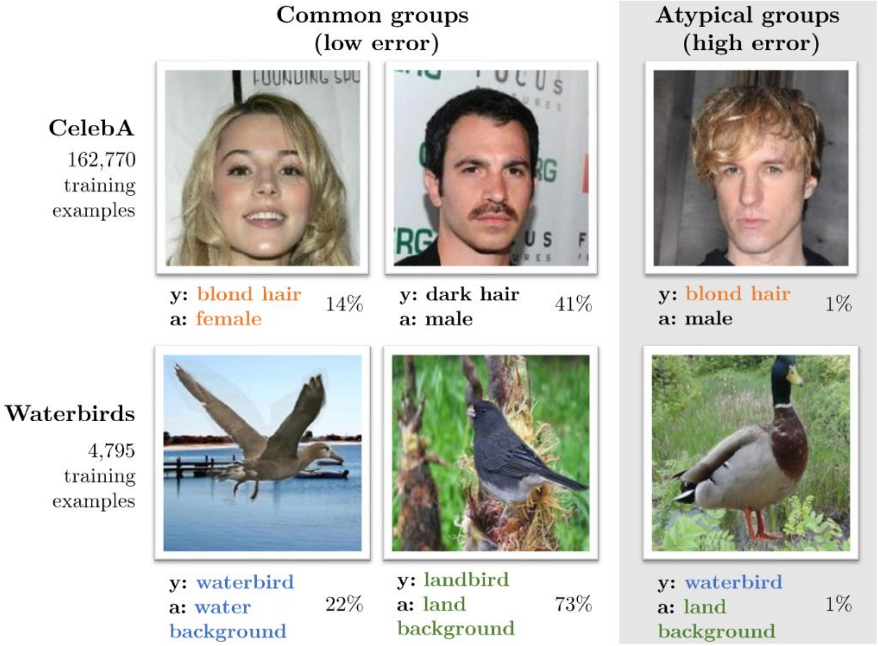
$$
\begin{aligned}
M(t) &= M_0 e^{-\eta(v+\Delta^{mix})t} + M^\infty(1 - e^{-\eta(v+\Delta^{mix})t}),\\
R_\pm(t) &= R^0_\pm e^{-\eta\Delta^{mix}t} + R^\infty_\pm(1 - e^{-\eta\Delta^{mix}t}) + k_1^\pm(e^{-\eta\Delta^{mix}t} - e^{-\eta(v+\Delta^{mix})t}),\\
Q(t) &= Q_0 e^{-\eta(2\Delta^{mix}-\eta\Delta^{2mix})t} + Q^\infty(1 - e^{-\eta(2\Delta^{mix}-\eta\Delta^{2mix})t})\\
&\quad + k_2(e^{-t(2\Delta^{mix}-\eta\Delta^{2mix})\eta} - e^{-t\Delta^{mix}\eta}) + k_3(e^{-t(2\Delta^{mix}-\eta\Delta^{2mix})\eta} - e^{-t(v+\Delta^{mix})\eta})\\
&\quad + k_4(e^{-t(2\Delta^{mix}-\eta\Delta^{2mix})\eta} - e^{-t(2v+2\Delta^{mix})\eta}),
\end{aligned}
$$

with $\Delta^{\text{mix}} = \rho\Delta_+ + (1-\rho)\Delta_-$ $\quad \Delta^{2\text{mix}} = \rho\Delta_+^2 + (1-\rho)\Delta_-^2$

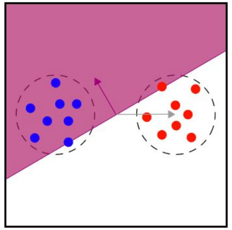
They characterise three timescales

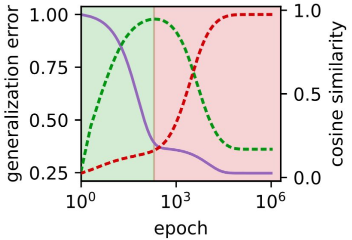
$$\tau_M = 1/[\eta(v + \Delta^{mix})] \qquad \tau_R = 1/(\eta\Delta^{mix}) \qquad \tau_Q = 1/[\eta(2\Delta^{mix} - \eta\Delta^{2mix})]$$

$\tau_R+\tau_M$: spurious correlations case



[Sagawa et al. 2020]



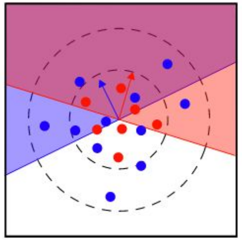


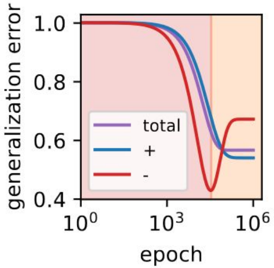
$$\tau_R = 1/(\eta\Delta^{mix})$$

$$\tau_M = 1/[\eta(v + \Delta^{mix})]$$

Spurious features are faster to learn but asymptotically disappear

$\tau_R + \tau_Q$: fairness case (centered)





Emergence of a bias crossing phenomenon

$$\tau_R = 1/(\eta\Delta^{mix}) \qquad \tau_Q = 1/[\eta(2\Delta^{mix} - \eta\Delta^{2mix})]$$

What's going on?

Initial dynamics

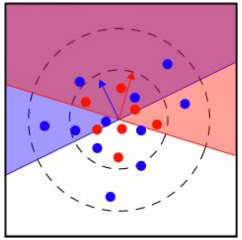
$$\left.\frac{d\epsilon_{g+}}{dt}\right|_{t=0} = -\eta^2\Delta^{mix}\Delta_+\left(\sqrt{\frac{2}{\pi\Delta_+}}\frac{R_+^\infty}{\eta} - 1\right)$$

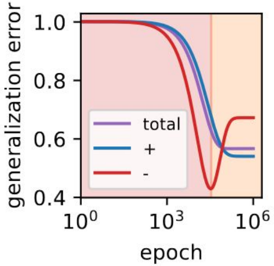
$$R_+^\infty = \sqrt{\frac{2}{\pi}}\frac{\rho\sqrt{\Delta_+} + T_\pm(1-\rho)\sqrt{\Delta_-}}{\Delta^{mix}}$$

$$\Rightarrow$$

$$T_\pm\sqrt{\frac{\Delta_+}{\Delta_-}} \leq \frac{d\epsilon_{g+}/dt|_{t=0}}{d\epsilon_{g-}/dt|_{t=0}} \leq \frac{1}{T_\pm}\sqrt{\frac{\Delta_+}{\Delta_-}}$$

$\tau_R + \tau_Q$: fairness case (centered)





Emergence of a bias crossing phenomenon

$$\tau_R = 1/(\eta\Delta^{mix}) \qquad \tau_Q = 1/[\eta(2\Delta^{mix} - \eta\Delta^{2mix})]$$

What's going on?

Initial dynamics ⟶ Saliency dominates

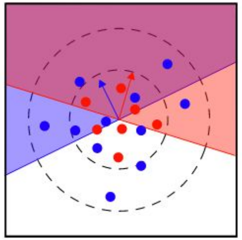
Asymptotic dynamics

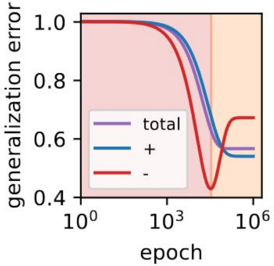
$$R_+^\infty = \sqrt{\frac{2}{\pi}}\frac{\rho\sqrt{\Delta_+} + T_\pm(1-\rho)\sqrt{\Delta_-}}{\Delta^{mix}}$$
$$R_-^\infty = \sqrt{\frac{2}{\pi}}\frac{T_\pm\rho\sqrt{\Delta_+} + (1-\rho)\sqrt{\Delta_-}}{\Delta^{mix}}$$

⟹

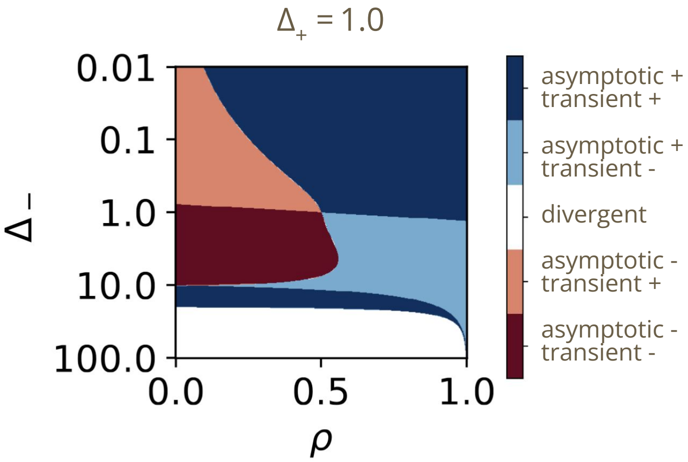
$$\rho\sqrt{\Delta_+} > (1-\rho)\sqrt{\Delta_-}$$

$\tau_R+\tau_Q$: fairness case (centered)





Emergence of a bias crossing phenomenon



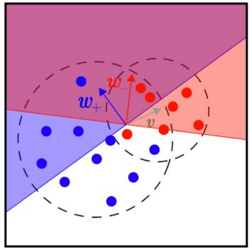
$$\tau_R = 1/(\eta\Delta^{mix}) \qquad \tau_Q = 1/[\eta(2\Delta^{mix} - \eta\Delta^{2mix})]$$

What's going on?

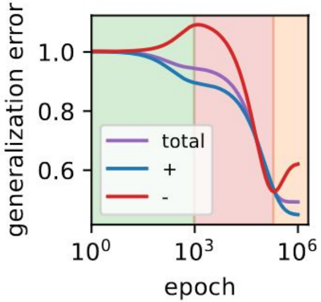
Initial dynamics $\longrightarrow$ Saliency dominates

Asymptotic dynamics $\longrightarrow$ Relative representation enters into play

$\tau_M + \tau_R + \tau_Q$: fairness case (general)



$$\tau_M = 1/[\eta(v + \Delta^{mix})]$$
$$\tau_R = 1/(\eta\Delta^{mix})$$
$$\tau_Q = 1/[\eta(2\Delta^{mix} - \eta\Delta^{2mix})]$$



Three phases

1. Green phase is driven by spurious correlation where the positive cluster is advantaged since it has greater representation and class imbalance.
2. Red phase is driven by greater variance where the negative cluster is learnt faster.
3. Orange phase where the student starts aligning with the positive taking into account relative representation.

Motivation

Understanding bias as a physicist

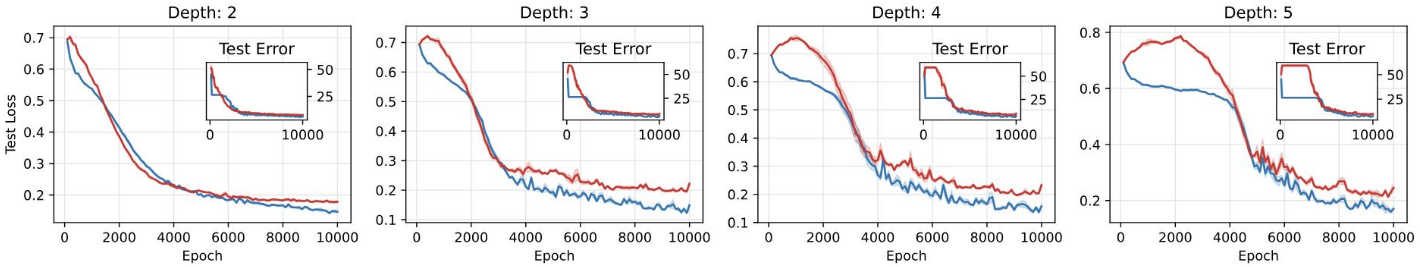
The dynamics of bias

Testing the results in the wild-ish

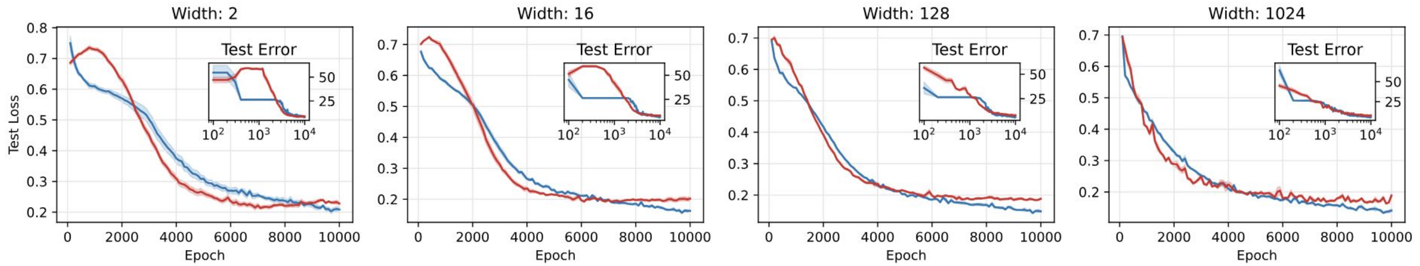
Conclusions

Numerical Experiments on Synthetic Data

Deeper...



...and wider ReLU networks in the synthetic framework

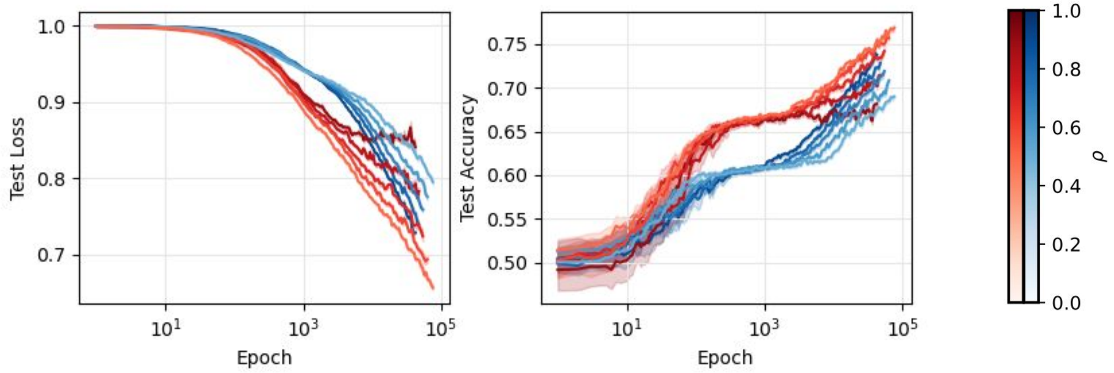


Numerical Experiments on CIFAR

3. Orange phase where the student starts aligning with the positive taking into account relative representation.

Let's revisit our starting experiment [**Bell & Sagun 2022**]:

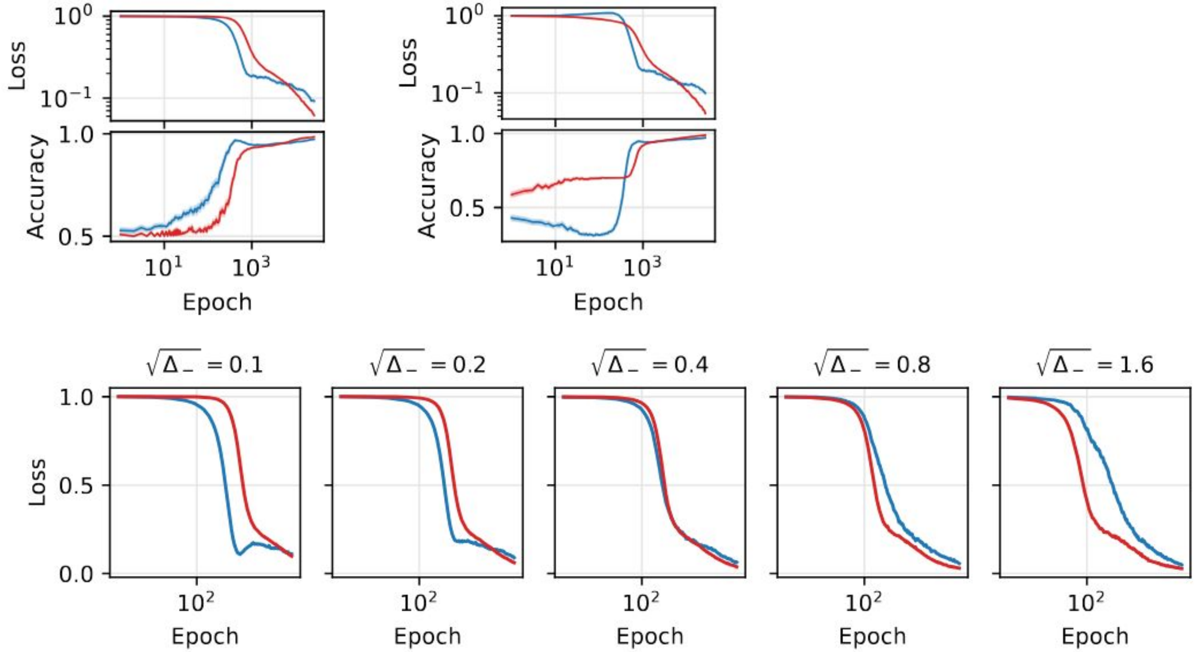
- a 2-layer NN trained online on CIFAR10 with MSE,
- divide the dataset in two populations and assign labels +1 and -1:
  - group 1: ['deer', 'bird', 'frog', 'horse']
  - group 2: ['cat', 'airplane', 'automobile', 'truck']
- Put the dataset back together with a mixture of ρ and (1-ρ) of group 1 and group 2 respectively.



Numerical Experiments on MNIST

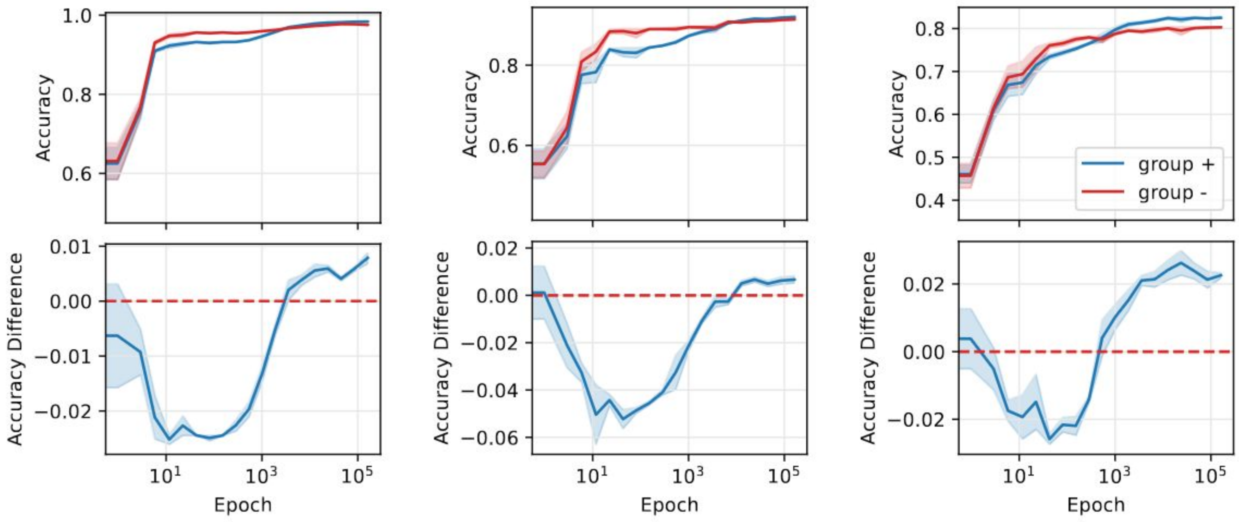
2. Red phase is driven by greater variance where the negative cluster is learnt faster.

Rotated MNIST



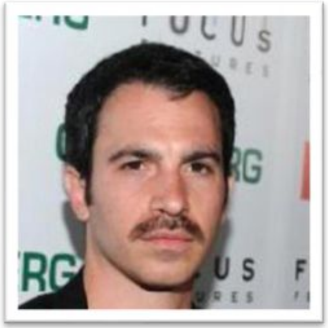
Numerical Experiments on CelebA

CelebA [Liu, et al. 2018]



(a) (Eye glass, Bags under eyes) (b) (Bangs, Blurry) (c) (Young, Blond Hair)

target group



Motivation

Understanding bias as a physicist

The dynamics of bias

Testing the results in the wild-ish

Conclusions

Conclusions

- ☆ Several aspects of the ML pipeline can potentially generate and amplify bias.
  - ★ Many of these are still underexplored!
  - ★ The assumptions behind our methods may fail.
- ☆ We focused on two aspects:
  - ★ The statistical properties of the data,
  - ★ The SGD learning dynamics.
- ☆ Our results show:
  - ★ The dynamics of bias can be non-monotonic with consequences on learning heuristics,
  - ★ We can characterise the features that attract the dynamics at different stages of learning.

Several open questions

- ☆ What is the effect of memorisation?
- ☆ What happens in multiclass settings when classes share similar structure?
- ☆ How prior knowledge (pre-training/continual learning) change this picture?
- ☆ ...

Aristide Baratin

Federica Gerace

Anchit Jain

Rozhin Nobahari

Negar Rostamzadeh

Luca Saglietti

Thank you! :D

References

★ Sarao Mannelli, Gerace, Rostamzadeh, Saglietti, PRE;

★ Jain, Nobahari, Baratin, Sarao Mannelli, NeurIPS 2024.

Support

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

stefSMlab.github.io =

CHALMERS UNIVERSITY OF TECHNOLOGY

GÖTEBORG UNIVERSITY

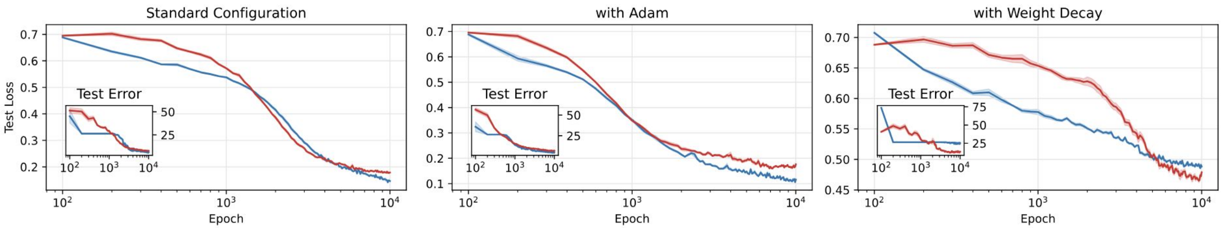


Figure 10: **Synthetic Data Simulation with alternate Training Protocols** We observe the 'double-crossing' phenomena in not only the loss curves, but also the error curves for the positive sub-population (blue) and the negative sub-population (red) *(left)*. The shaded areas quantify the standard deviation obtained across 10 seeds. We observe similar behavior when using Adam *(middle)* and weight decay *(right)*. The data distribution parameters are $d = 100, v = 4, \rho = 0.7, \Delta_+ = 0.1, \Delta_- = 1, T_\pm = 0.9, \eta = 0.01, \alpha_+ = 0.471, \alpha_- = -0.188$

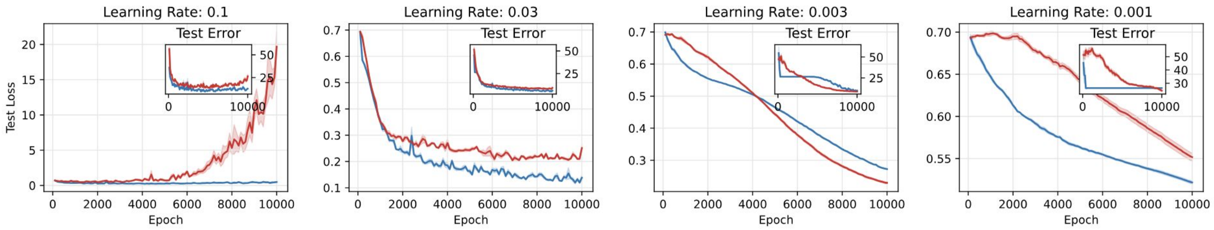


Figure 13: **Ablations across Learning Rates** Larger learning rates can lead to instability *(left)*. If training is stable however, we observe the 'crossing' phenomena as usual, just at different time scales due to different speeds of training.