



Announcements

- ▶ A1 due Tuesday
- ▶ A2 released Tuesday
- ▶ Fairness response (in class today) due in 1 week
- ▶ Seating chart



Recap: Multiclass

- ▶ Different weights: $\operatorname{argmax}_{y \in \mathcal{Y}} w_y^\top f(x)$
 - ▶ Generalizes to neural networks: $f(x)$ is the first $n-1$ layers of the network, then you multiply by a final linear layer at the end
- ▶ Different features: $\operatorname{argmax}_{y \in \mathcal{Y}} w^\top f(x, y)$
 - ▶ Suppose $\mathcal{Y}$ is a structured label space (part-of-speech tags for each word in a sentence). $f(x, y)$ looks at each POS tag individually, no easy way to conceptualize what w_y is for different weights
- ▶ For linear multiclass classification with discrete classes, these are identical



Recap: Multiclass Logistic Regression

$$P(y = \hat{y} | \bar{x}) = \frac{\exp(\bar{w}_{\hat{y}}^\top f(\bar{x}))}{\sum_{y \in \mathcal{Y}} \exp(\bar{w}_{y'}^\top f(\bar{x}))}$$

- Update: let $y^{(i)}$ be the gold label

$$\bar{w}_{y^{(i)}} \leftarrow \bar{w}_{y^{(i)}} + \alpha f(\bar{x}^{(i)}) \left(1 - P \left(y = y^{(i)} \mid \bar{x}^{(i)} \right) \right)$$

For all other y'

$$\bar{w}_{y'} \leftarrow \bar{w}_{y'} - \alpha f(\bar{x}^{(i)}) P \left(y = y^{(i)} \mid \bar{x}^{(i)} \right)$$



Recap: Multiclass Logistic Regression



Today

- ▶ Multiclass examples
- ▶ Fairness in classification
- ▶ Intro to neural networks

Multiclass Examples



Text Classification

A Cancer Conundrum: Too Many Drug Trials, Too Few Patients

Breakthroughs in immunotherapy and a rush to develop profitable new treatments have brought a crush of clinical trials scrambling for patients.

By GINA KOLATA



→ Health

Yankees and Mets Are on Opposite Tracks This Subway Series

As they meet for a four-game series, the Yankees are playing for a postseason spot, and the most the Mets can hope for is to play spoiler.

By FILIP BONDY



→ Sports

~20 classes

- 20 Newsgroups, Reuters, Yahoo! Answers, ...



Entailment

- ▶ Three-class task over sentence pairs
- ▶ Not clear how to do this with simple bag-of-words features

A soccer game with multiple males playing.

ENTAILS

Some men are playing a sport.

A black race car starts up in front of a crowd of people.

CONTRADICTS

A man is driving down a lonely road

A smiling costumed woman is holding an umbrella.

NEUTRAL

A happy woman in a fairy costume holds an umbrella.



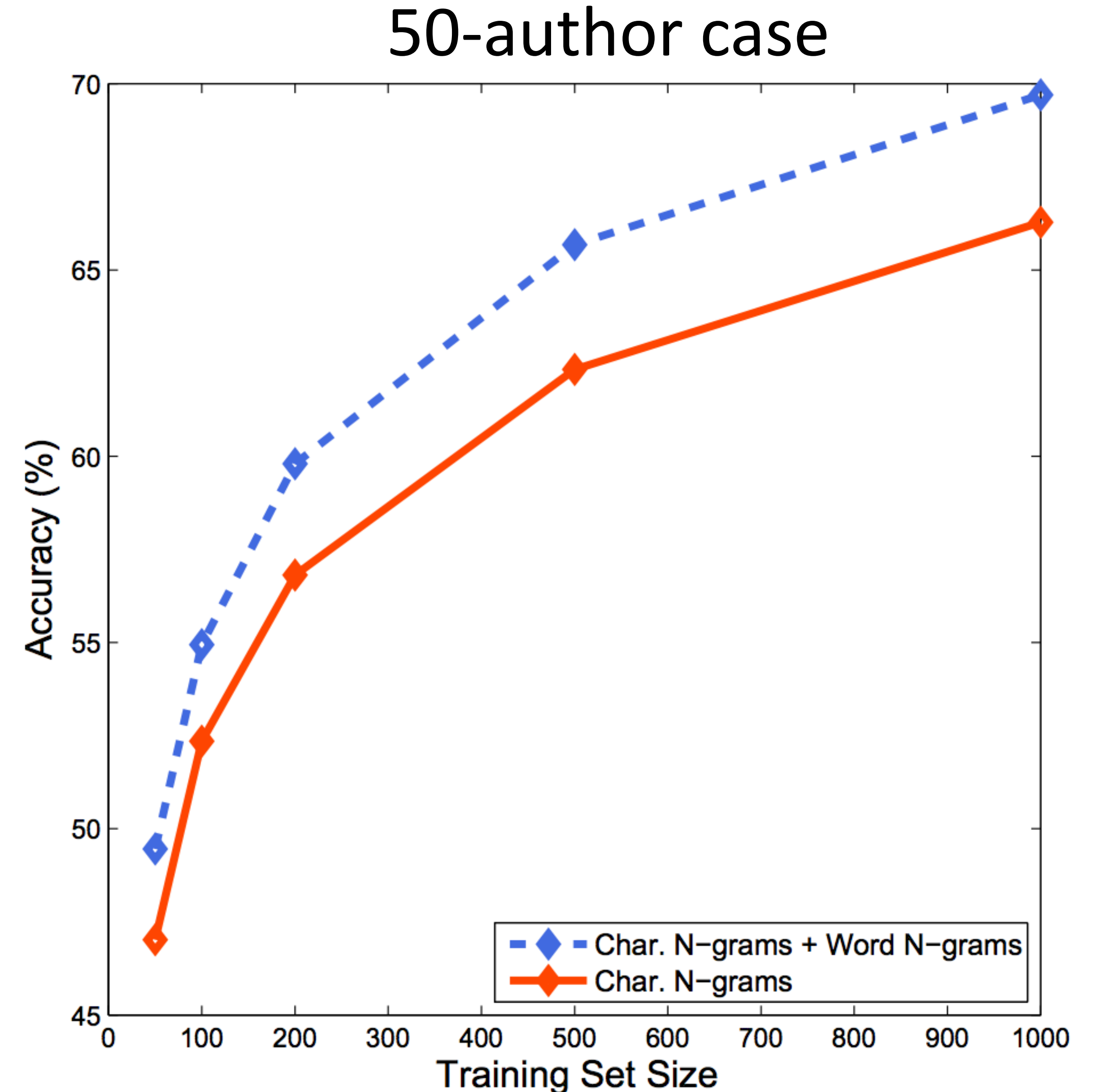
Authorship Attribution

- ▶ Statistical methods date back to 1930s and 1940s
 - ▶ Based on handcrafted heuristics like stopword frequencies
 - ▶ Early work: Shakespeare's plays, Federalist papers (Hamilton v. Madison)
- ▶ Twitter: given a bunch of tweets, can we figure out who wrote them?
 - ▶ Schwartz et al. EMNLP 2013: 500M tweets, take 1000 users with at least 1000 tweets each
- ▶ Task: given a held-out tweet by one of the 1000 authors, who wrote it?



Authorship Attribution

- ▶ SVM with character 4-grams, words 2-grams through 5-grams
- ▶ 1000 authors, 200 tweets per author => 30% accuracy
- ▶ 50 authors, 200 tweets per author => 71.2% accuracy





Authorship Attribution

- ▶ k-signature: n-gram that appears in k% of the authors tweets but not appearing for anyone else — suggests why these are so effective

Signature Type	10%-signature	Examples
Character n-grams	‘ ^ _ ^ ’	REF oh ok <u>^ _ ^</u> Glad you found it!
		Hope everyone is having a good afternoon <u>^ _ ^</u>
		REF Smirnoff lol keeping the goose in the freezer <u>^ _ ^</u>
	‘yew ’	gurl <u>yew</u> serving me tea nooch
		REF about wen <u>yew</u> and ronnie see each other
		REF lol so <u>yew</u> goin to check out tini’s tonight huh???

Fairness



Fairness in Classification

- ▶ Classifiers can be used to make real-world decisions:
 - ▶ Who gets an interview?
 - ▶ Who should we lend money to?
 - ▶ Is this online activity suspicious?
 - ▶ Is a convicted person likely to re-offend?
- ▶ Humans making these decisions are typically subject to anti-discrimination laws; how do we ensure classifiers are *fair* in the same way?
- ▶ Many other factors to consider when deploying classifiers in the real world (e.g., impact of a false positive vs. a false negative) but we'll focus on fairness here



Fairness Response (SUBMIT ON CANVAS)

Consider having each data instance x associated with a **protected attribute A** when making a prediction. For example, suppose for sentiment analysis we also had information about the **ethnicity of the director** of the movie being reviewed.

- ▶ What do **you** think it would mean for a classification model to be discriminatory in this context? Try to be as precise as you can!
- ▶ Suppose we add A as an additional “word” to each example, so our bag-of-words can use it as part of the input. Do you think the unigram model might be discriminatory according to your criterion? Why or why not?
- ▶ Suppose we ignore A (use our existing model). Do you think the unigram model might be discriminatory according to your criterion above? Why or why not?
- ▶ Suppose we enforce that the model must predict at least $k\%$ positives across every value of A ; that is, if you filter to only the data around a particular ethnicity, the model must predict at least $k\%$ positives on that data slice. Is this fair? Why/why not?

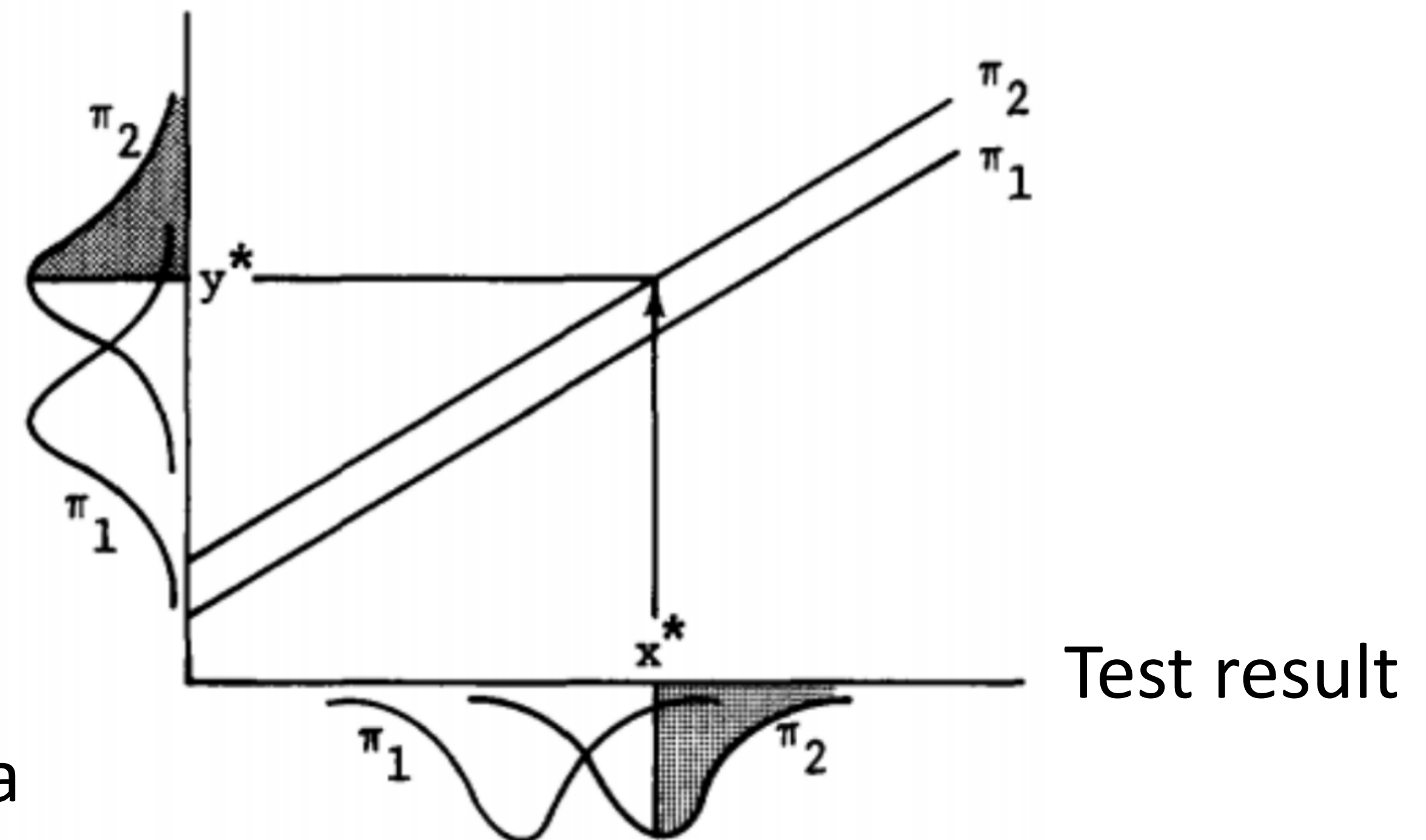


Fairness in Classification

Idea 1: Classifiers need to be evaluated beyond just accuracy

- ▶ T. Anne Cleary (1966-1968): a test is biased if prediction on a subgroup makes *consistent* nonzero prediction errors compared to the aggregate
- ▶ Individuals of X group could still score lower on average. But the *errors* should not be consistently impacting X
- ▶ Member of π_1 has a test result higher than a member of π_2 for the same ground truth ability. Test penalizes π_2

Ground truth





Fairness in Classification

Idea 1: Classifiers need to be evaluated beyond just accuracy

- ▶ Thorndike (1971), Petersen and Novik (1976): fairness in classification: ratio of predicted positives to ground truth positives must be approximately the same for each group (“equalized odds”)
 - ▶ Group 1: 50% positive movie reviews. Group 2: 60% positive movie reviews
 - ▶ A classifier classifying 50% positive in both groups is unfair, regardless of accuracy
- ▶ Allows for different criteria across groups: imposing different classification thresholds actually can give a fairer result
- ▶ There are many other criteria we could use as well — this isn’t the only one!
- ▶ Can’t we just make our classifiers not depend on sensitive features like gender?

Petersen and Novik (1976)
Hutchinson and Mitchell (2018)



Discrimination

Idea 2: It is easy to build classifiers that discriminate even *without meaning to*

- ▶ A feature might correlate with minority group X and penalize that group:
 - ▶ Bag-of-words features can identify dialects of English like AAVE or code-switching (using two languages). **Could it identify movie titles in other languages?**
 - ▶ ZIP code as a feature is correlated with race
- ▶ Reuters: “Amazon scraps secret AI recruiting tool that showed bias against women”
 - ▶ “Women’s X” organization, women’s colleges were negative-weight features
 - ▶ Accuracy will not catch these problems, very complex to evaluate depending on what humans did in the **actual** recruiting process

Credit: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>



Takeaways

- ▶ What minority groups in the population should I be mindful of? (Review sentiment: movies with female directors, foreign films, ...)
- ▶ Can I check one of these fairness criteria?
- ▶ Do aspects of my system or features it uses introduce potential correlations with protected classes or minority groups?

Neural Networks



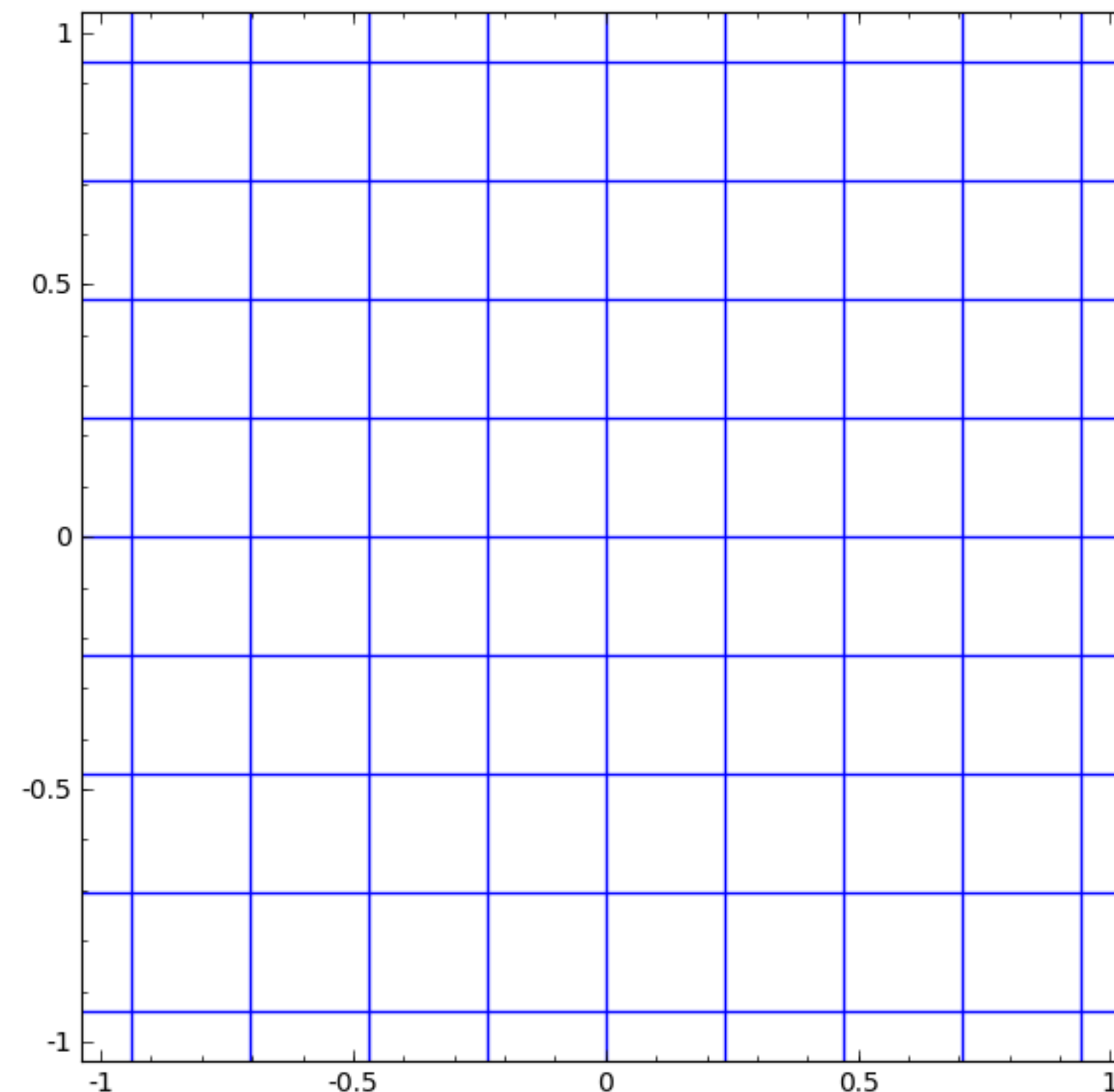
Neural Networks

$$\mathbf{z} = g(Vf(\mathbf{x}) + \mathbf{b})$$

Nonlinear transformation Warp space Shift

$$y_{\text{pred}} = \operatorname{argmax}_y \mathbf{w}_y^\top \mathbf{z}$$

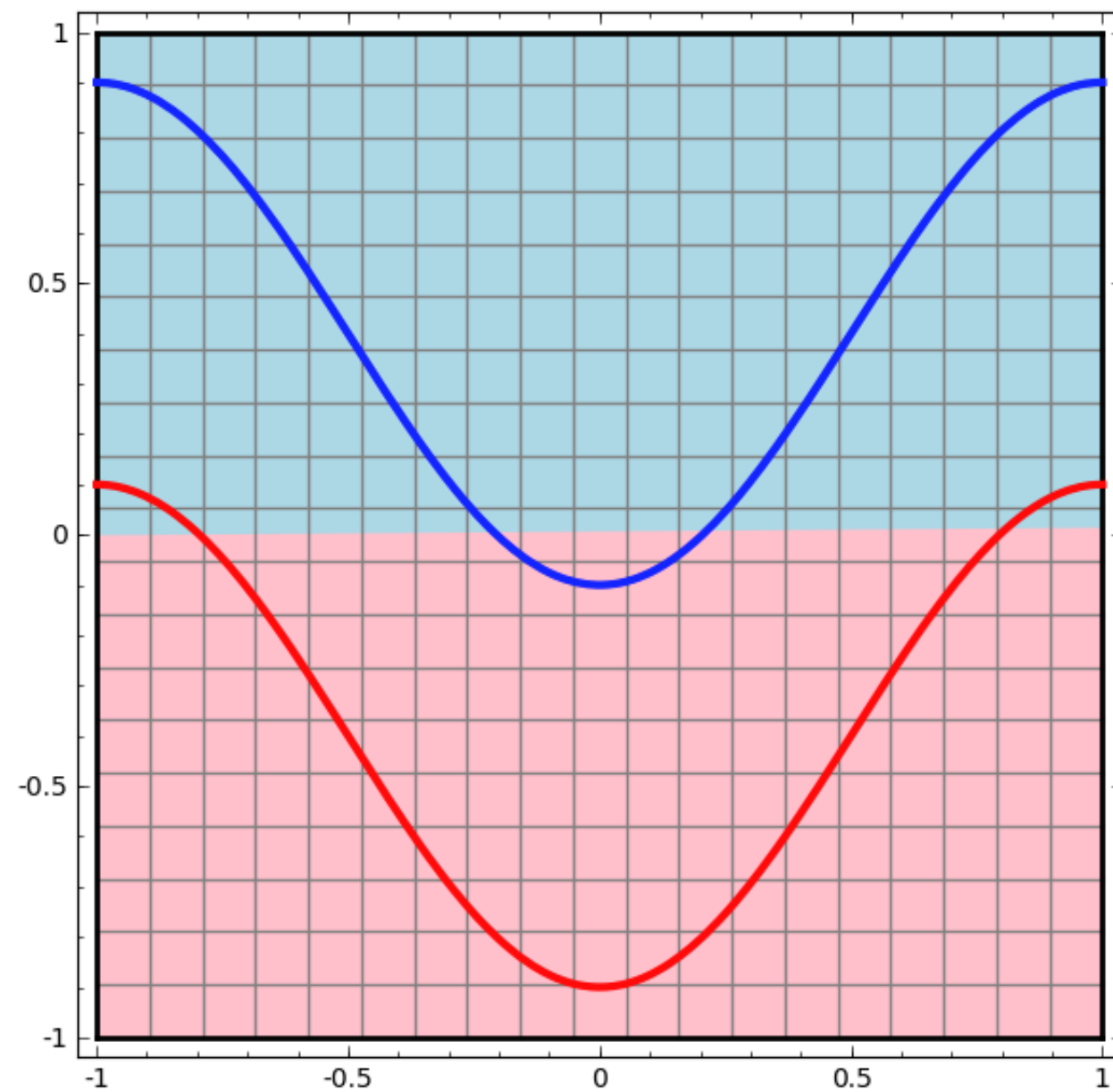
- Ignore shift / $+\mathbf{b}$ term for the rest of the course



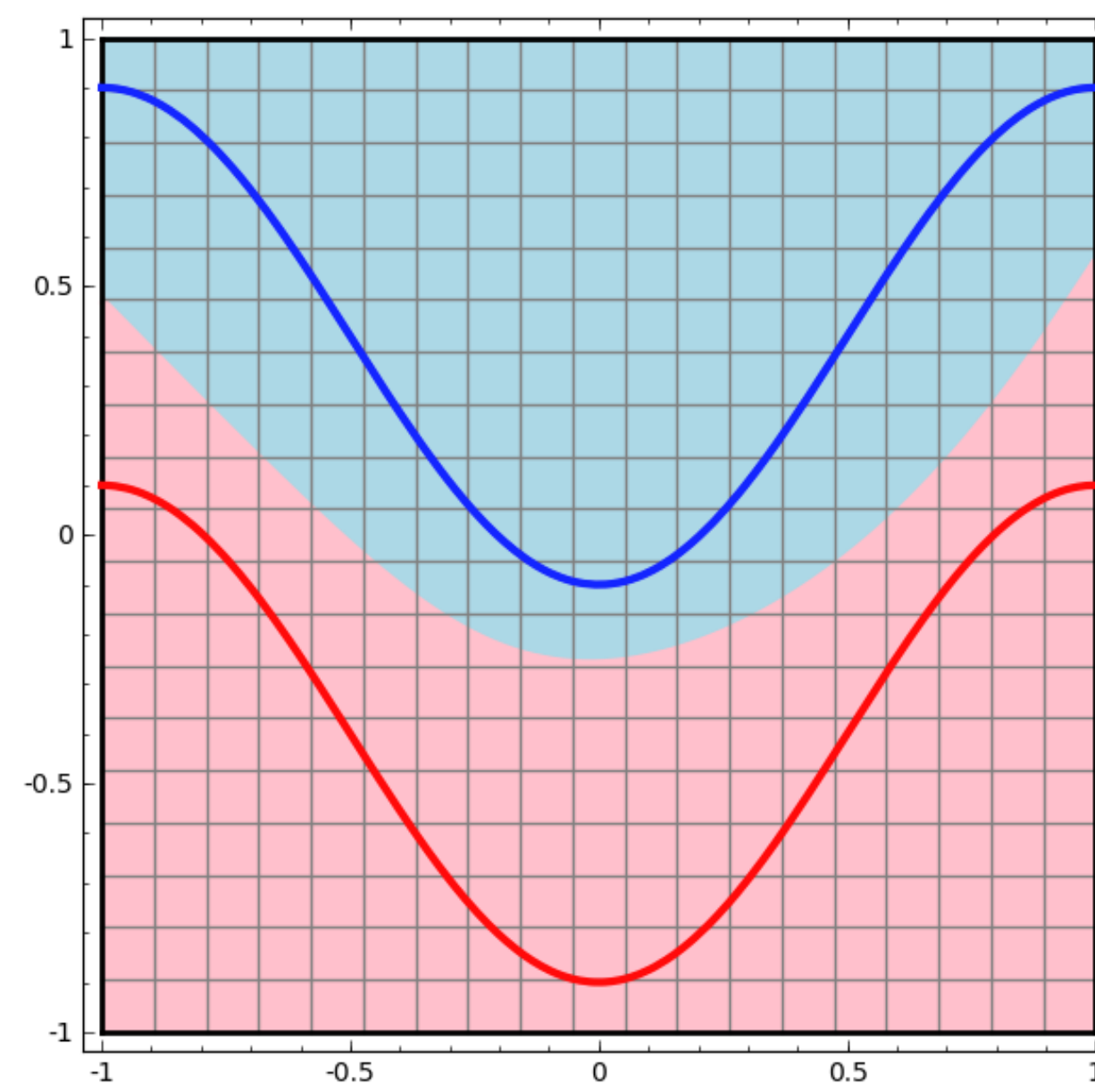


Neural Networks

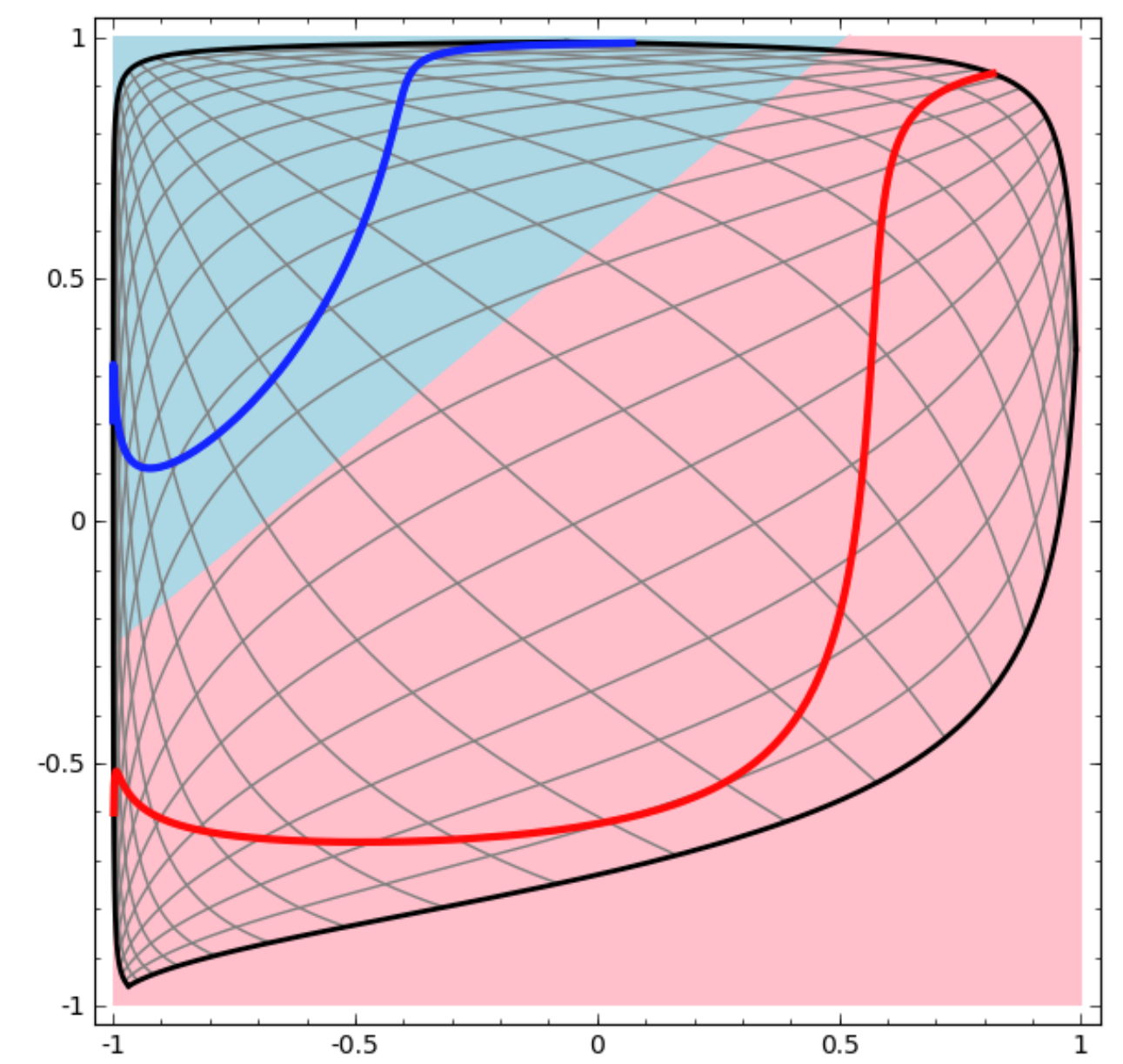
Linear classifier



Neural network



Linear classification
in the transformed
space!





Deep Neural Networks

$$\mathbf{z}_1 = g(V_1 f(\mathbf{x}))$$

$$\mathbf{z}_2 = g(V_2 \mathbf{z}_1)$$

...

$$y_{\text{pred}} = \operatorname{argmax}_y \mathbf{w}_y^\top \mathbf{z}_n$$

