Security and Fairness of Deep Learning

Feature-wise Bias Amplification

Klas Leino

CMU

Spring 2019

Overview

- Background
    - Motivation
    - Bias Amplification
    - Hypothesis
- Analysis
    - Formalization
    - Naïve Bayes case
    - Feature Asymmetry
    - Feature-wise bias amplification effect
- Mitigation
    - Feature parity & experts
    - Evaluation

Original Motivation

- Zhao et al. [1] show found that a model trained to predict actions and agents in images predicted that the agent in cooking scenes was a woman at a higher rate than in the training data
- This phenomenon is called *bias amplification*
- Setting of Zhao et al. is slightly different from our work [2], because they condition on a *protected attribute*, but our hypothesized mechanism is conceptually similar

[1] Zhao et al. "Men also Like Shopping: Reducing Gender Bias Amplification Using Corpus-level Constraints" ArXiv

[2] Leino et al. "Feature-wise Bias Amplification" ArXiv

Definition – Prior Distribution of Data

- Consider a dataset with 60% class “A” and 40% class “B”
  - We say the *prior distribution of the data* is (0.6, 0.4)

- Suppose a model labels 55% of its predictions “A” and 45% “B”
  - We say the *prior distribution of the predictions* is (0.55, 0.45)

Definition – Bias Amplification

We say a model exhibits *bias amplification* if the prior distribution of the model’s predictions does not match that of the data.

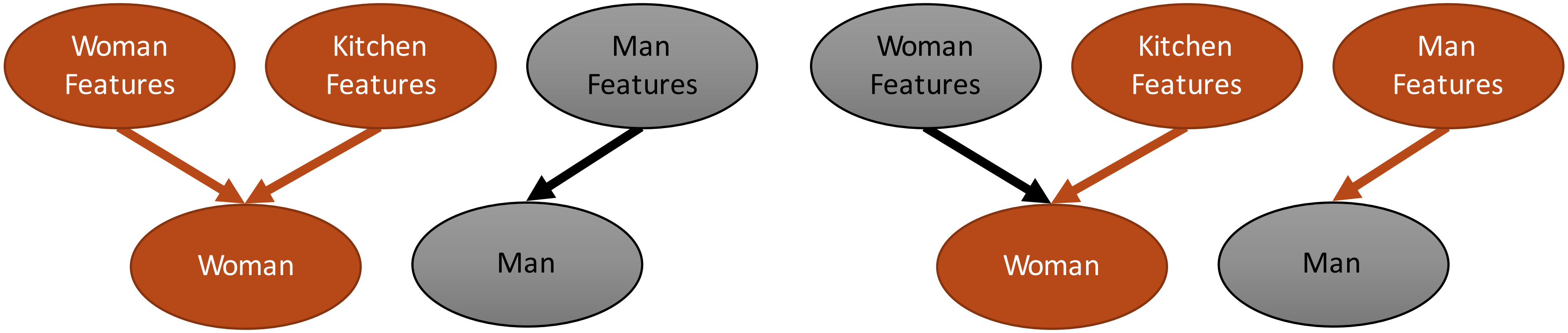
In particular, we don’t want the prior distribution of the predictions to be more disparate than that of the data.

Example

Class	Man	Woman
Data prior	33%	67%
Pred. prior	16%	84%

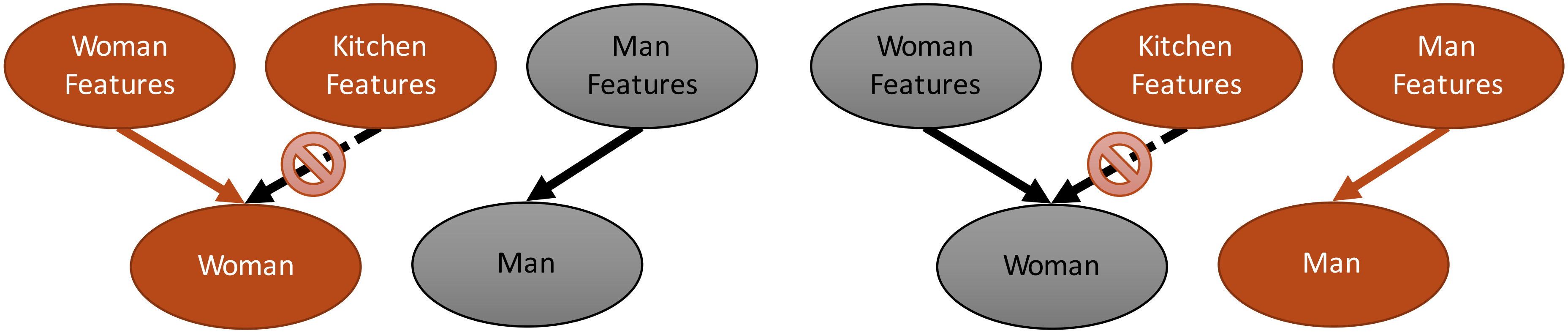
Hypothesis

- Model learns to detect *woman* features, *man* features, and *kitchen* features
- Kitchen features contribute to *woman* class due to correlation in training data



Hypothesis

- Idea: distinguish between woman features and kitchen features, and prevent kitchen features from contributing to woman class



Recall – Experts

- An *expert* is a sub-model which has increased accuracy on a particular class from the original model
- Experts exploit the fact that performance can often be improved via the removal of “distracting” features

Encouraging Preliminary Results

Class	0	1	2	3	4	5	6	7	8	9
Data prior	20%	5%	5%	10%	10%	10%	10%	10%	10%	10%
Original pred. 79.6% Accuracy	19.6%	4.0%	3.3%	10.5%	10.2%	9.2%	11.1%	10.2%	11.2%	10.8%
Compressed pred. 79.8% Accuracy	20.0%	4.6%	5.0%	10.1%	10.1%	10.0%	9.9%	9.9%	10.3%	10.2%
Data prior	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Original pred. 76.9% Accuracy	11.6%	7.8%	5.9%	10.8%	10.5%	9.4%	11.6%	10.2%	10.8%	11.3%
Compressed pred. 77.6% Accuracy	11.9%	8.6%	7.9%	10.4%	10.5%	10.2%	10.1%	9.8%	10.1%	10.5%

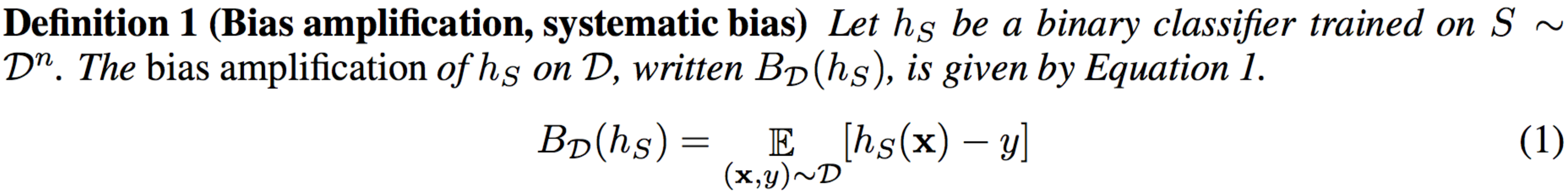
Overview

- Background
    - Motivation
    - Bias Amplification
    - Hypothesis
- Analysis
    - Formalization
    - Naïve Bayes case
    - Feature Asymmetry
    - Feature-wise bias amplification effect
- Mitigation
    - Feature parity & experts
    - Evaluation

Formalization

- D : distribution of labeled points
- S : training set obtained via n i.i.d. samples drawn from D
- h_S : binary classifier learned via some learning rule, R

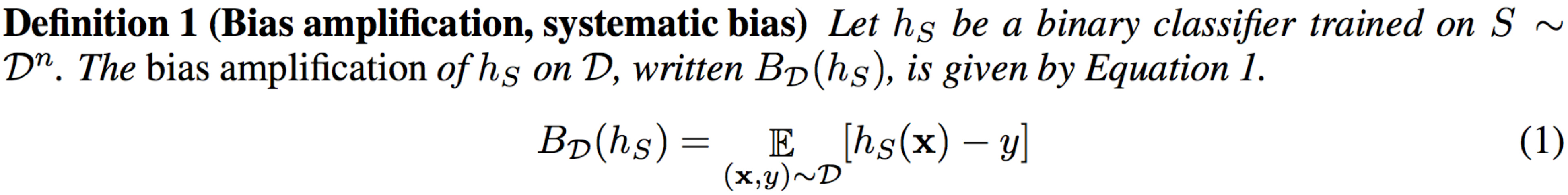
Bias Amplification



Definition 1 (Bias amplification, systematic bias) *Let* h_S *be a binary classifier trained on* $S \sim \mathcal{D}^n$. *The* bias amplification *of* h_S *on* $\mathcal{D}$, *written* $B_{\mathcal{D}}(h_S)$, *is given by Equation 1.*

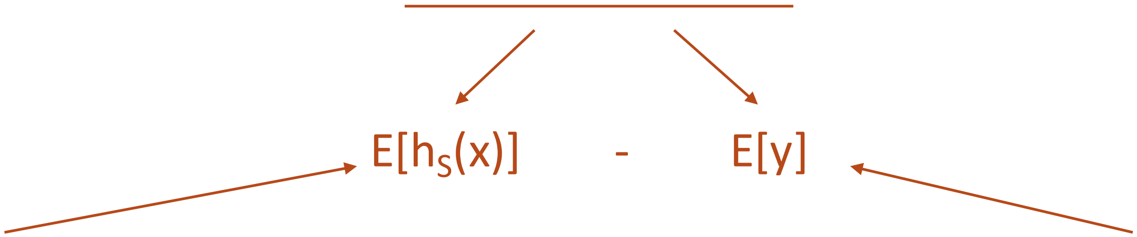
$$B_{\mathcal{D}}(h_S) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[h_S(\mathbf{x}) - y] \tag{1}$$

Bias Amplification



Definition 1 (Bias amplification, systematic bias) *Let h_S be a binary classifier trained on $S \sim \mathcal{D}^n$. The* bias amplification *of h_S on $\mathcal{D}$, written $B_{\mathcal{D}}(h_S)$, is given by Equation 1.*

$$B_{\mathcal{D}}(h_S) = \underset{(\mathbf{x},y)\sim\mathcal{D}}{\mathbb{E}}[h_S(\mathbf{x}) - y] \tag{1}$$

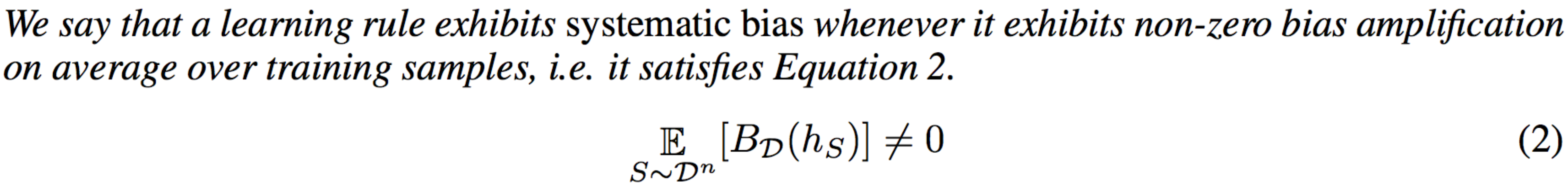


E[h_S(x)] - E[y]

Fraction of instances labeled “class 1” by the model

Fraction of instances with true label “class 1” in the data

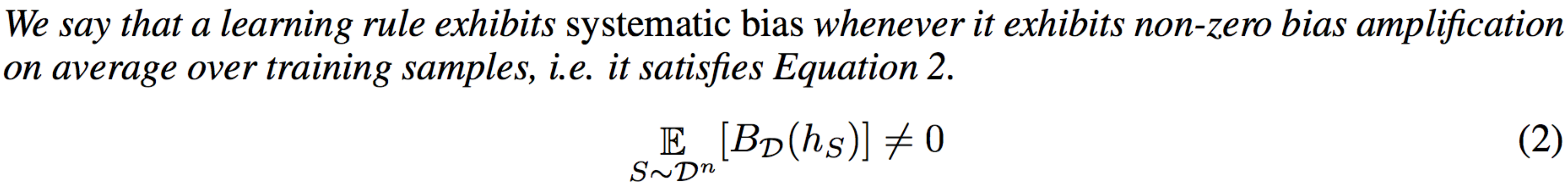
Systematic Bias



We say that a learning rule exhibits systematic bias *whenever it exhibits non-zero bias amplification on average over training samples, i.e. it satisfies Equation 2.*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}[B_{\mathcal{D}}(h_S)] \neq 0 \tag{2}$$

Systematic Bias

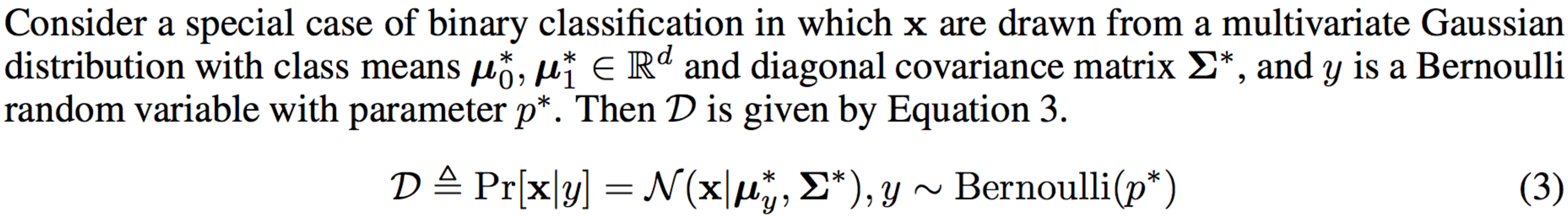


We say that a learning rule exhibits systematic bias *whenever it exhibits non-zero bias amplification on average over training samples, i.e. it satisfies Equation 2.*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^n} [B_{\mathcal{D}}(h_S)] \neq 0 \tag{2}$$

i.e., given a random training set, we expect to get non-zero bias. We include this expectation since we may get bias on occasion simply by being unlucky with the training set we chose

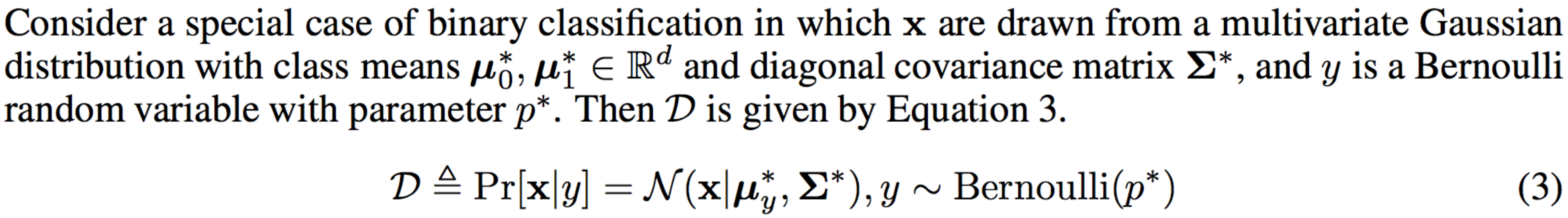
Simplified Formal Setting



Consider a special case of binary classification in which $\mathbf{x}$ are drawn from a multivariate Gaussian distribution with class means $\boldsymbol{\mu}_0^*, \boldsymbol{\mu}_1^* \in \mathbb{R}^d$ and diagonal covariance matrix $\boldsymbol{\Sigma}^*$, and y is a Bernoulli random variable with parameter p^*. Then $\mathcal{D}$ is given by Equation 3.

$$\mathcal{D} \triangleq \Pr[\mathbf{x}|y] = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_y^*, \boldsymbol{\Sigma}^*), y \sim \text{Bernoulli}(p^*) \tag{3}$$

Simplified Formal Setting

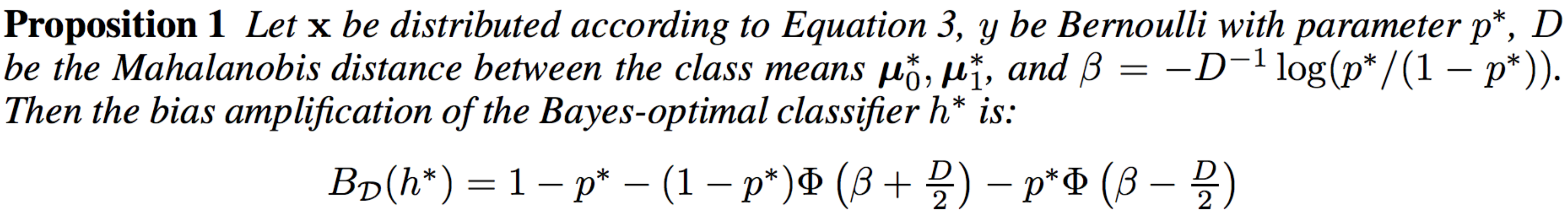


Consider a special case of binary classification in which $\mathbf{x}$ are drawn from a multivariate Gaussian distribution with class means $\boldsymbol{\mu}_0^*, \boldsymbol{\mu}_1^* \in \mathbb{R}^d$ and diagonal covariance matrix $\boldsymbol{\Sigma}^*$, and y is a Bernoulli random variable with parameter p^*. Then $\mathcal{D}$ is given by Equation 3.

$$\mathcal{D} \triangleq \Pr[\mathbf{x}|y] = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_y^*, \boldsymbol{\Sigma}^*), y \sim \text{Bernoulli}(p^*) \qquad (3)$$

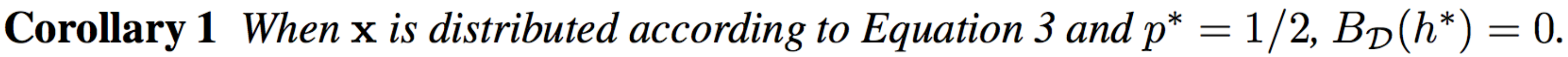
i.e., we assume our data is Gaussian and distributed according to the Naïve Bayes assumption (features are independent conditioned on the class). This means that the Naïve Bayes classifier is the optimal classifier for our data.

Simplified Formal Setting



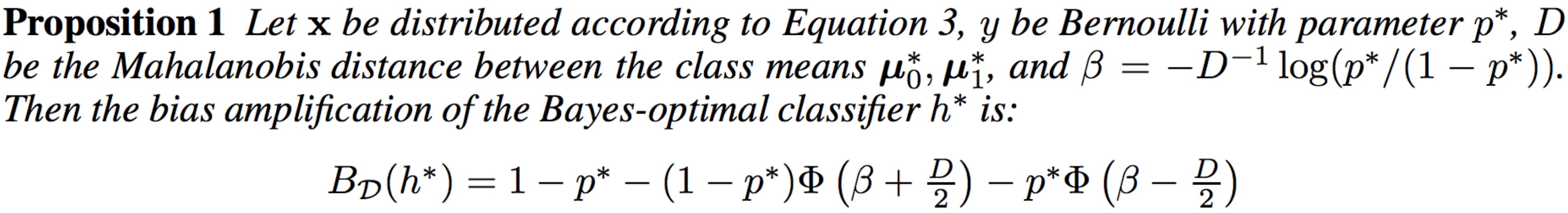
Proposition 1 *Let* $\mathbf{x}$ *be distributed according to Equation 3,* y *be Bernoulli with parameter* p^*, D *be the Mahalanobis distance between the class means* $\boldsymbol{\mu}_0^*, \boldsymbol{\mu}_1^*$, *and* $\beta = -D^{-1}\log(p^*/(1-p^*))$. *Then the bias amplification of the Bayes-optimal classifier* h^* *is:*

$$B_{\mathcal{D}}(h^*) = 1 - p^* - (1-p^*)\Phi\left(\beta + \tfrac{D}{2}\right) - p^*\Phi\left(\beta - \tfrac{D}{2}\right)$$



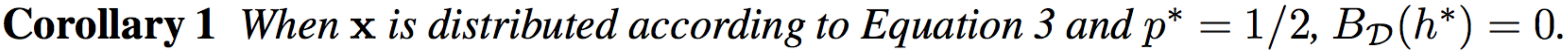
Corollary 1 *When* $\mathbf{x}$ *is distributed according to Equation 3 and* $p^* = 1/2$, $B_{\mathcal{D}}(h^*) = 0$.

Simplified Formal Setting



Proposition 1 *Let* $\mathbf{x}$ *be distributed according to Equation 3,* y *be Bernoulli with parameter* p^**,* D *be the Mahalanobis distance between the class means* $\boldsymbol{\mu}_0^*, \boldsymbol{\mu}_1^*$*, and* $\beta = -D^{-1}\log(p^*/(1-p^*))$*. Then the bias amplification of the Bayes-optimal classifier* h^* *is:*

$$B_{\mathcal{D}}(h^*) = 1 - p^* - (1-p^*)\Phi\left(\beta + \frac{D}{2}\right) - p^*\Phi\left(\beta - \frac{D}{2}\right)$$



Corollary 1 *When* $\mathbf{x}$ *is distributed according to Equation 3 and* $p^* = 1/2$*,* $B_{\mathcal{D}}(h^*) = 0$*.*

i.e., if the classes of Naïve Bayes data are equally distributed, the optimal classifier will *never* exhibit bias amplification

Bias Amplification using LR & SGD

- In practice, we would be more likely to use LR models than NB models, since they make fewer assumptions about the data
- For Naïve Bayes data (Equation 3) LR converges to Bayes-optimal *when given sufficient data*
- We show that prior to convergence, LR trained with SGD can exhibit systematic bias amplification *even on data for which the Bayes-optimal classifier would not*

Strong & Weak Features

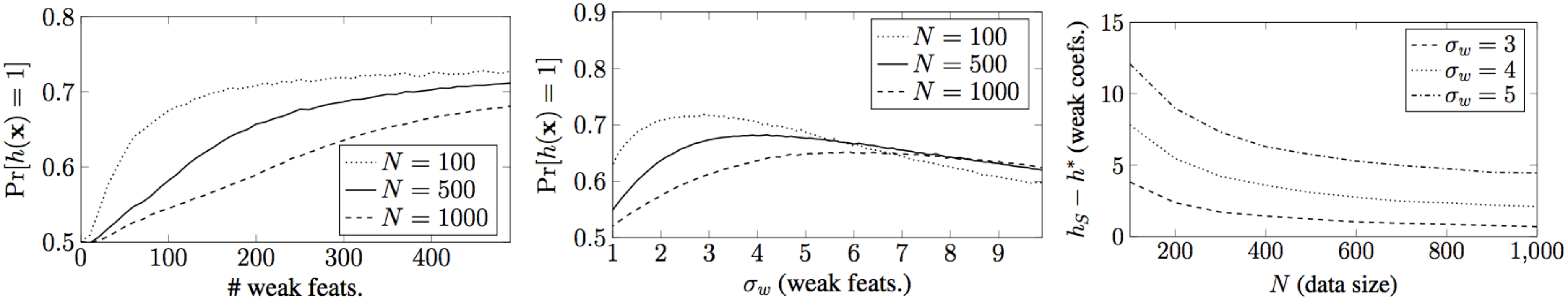
- At a high level, we consider “strong” features, which have high weight (high predictive power), and “weak” features, which have low weight (low predictive power)
- In practice, there is not a dichotomy, but we use a dichotomy for the sake of argument, and the general intuition holds for more nuanced cases

Direction of a Feature

We consider the *direction* of a feature to be the class for which the feature has positive weight. I.e., features are oriented towards whichever class they supply positive evidence for

Feature Asymmetry

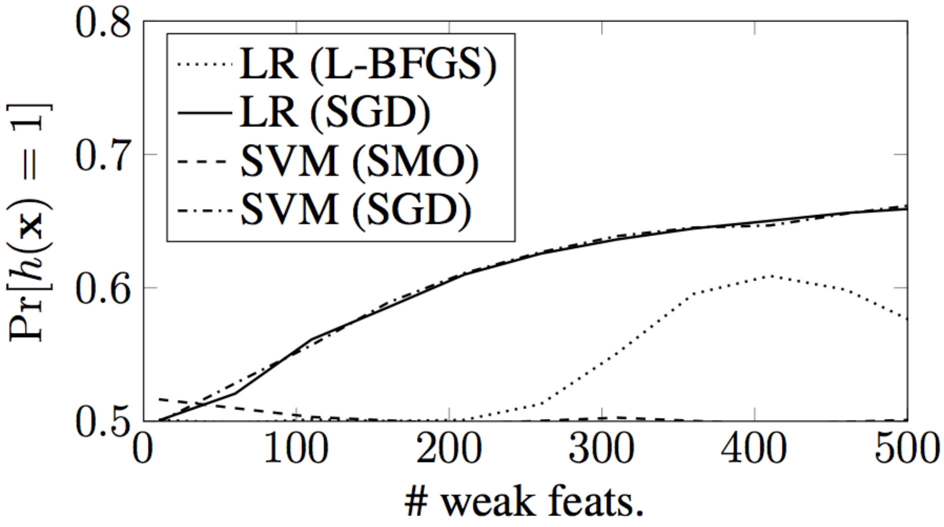
- SGD tends to overestimate the importance of weak features prior to convergence
- When there is an imbalance of weak features, the overestimation accumulates in favor of the class with more weak features



(a) (b) (c)

What is to Blame?

- Was this a consequence of the learning rule (SGD) or the target (LR)?
  - ➢Evidence shows effect is consistent across models trained with SGD, suggesting the effect is attributable to the learning rule



Some Intuition

SGD effectively applies an implicit L2 regularization. This spreads out the influence of each feature rather than letting the influence go directly to the most important features. This leads weak features to be overestimated and strong features to be overestimated.

**Note that these claims are all intuitive rather than rigorous, so take them with a grain of salt*

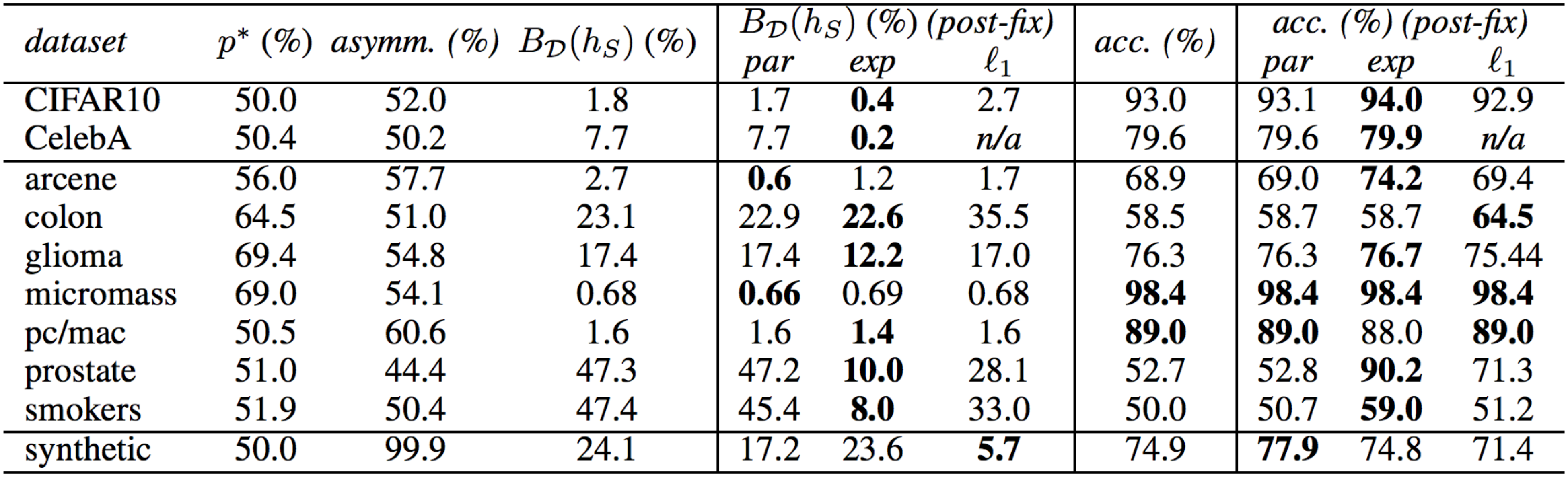
Overview

- Background
  - Motivation
  - Bias Amplification
  - Hypothesis
- Analysis
  - Formalization
  - Naïve Bayes case
  - Feature Asymmetry
  - Feature-wise bias amplification effect
- Mitigation
  - Feature parity & experts
  - Evaluation

Possible Fixes

- Feature parity
  - Eliminate features such that the number of features in each direction is equal, removing lowest-weight features first
  - Generally too weak of an impact to make a large difference
- Experts
  - Select the expert that minimizes bias amplification subject to not harming accuracy
- L1 regularization
  - Encourages sparsity which might eliminate weak features
  - Tends not to be sufficient for desired effect

Results



dataset	p^* (%)	*asymm.* (%)	$B_{\mathcal{D}}(h_S)$ (%)	$B_{\mathcal{D}}(h_S)$ (%) *(post-fix)*			*acc.* (%)	*acc.* (%) *(post-fix)*		
				par	*exp*	ℓ_1		*par*	*exp*	ℓ_1
CIFAR10	50.0	52.0	1.8	1.7	**0.4**	2.7	93.0	93.1	**94.0**	92.9
CelebA	50.4	50.2	7.7	7.7	**0.2**	*n/a*	79.6	79.6	**79.9**	*n/a*
arcene	56.0	57.7	2.7	**0.6**	1.2	1.7	68.9	69.0	**74.2**	69.4
colon	64.5	51.0	23.1	22.9	**22.6**	35.5	58.5	58.7	58.7	**64.5**
glioma	69.4	54.8	17.4	17.4	**12.2**	17.0	76.3	76.3	**76.7**	75.44
micromass	69.0	54.1	0.68	**0.66**	0.69	0.68	**98.4**	**98.4**	**98.4**	**98.4**
pc/mac	50.5	60.6	1.6	1.6	**1.4**	1.6	**89.0**	**89.0**	88.0	**89.0**
prostate	51.0	44.4	47.3	47.2	**10.0**	28.1	52.7	52.8	**90.2**	71.3
smokers	51.9	50.4	47.4	45.4	**8.0**	33.0	50.0	50.7	**59.0**	51.2
synthetic	50.0	99.9	24.1	17.2	23.6	**5.7**	74.9	**77.9**	74.8	71.4

Some Intuition

- L2 regularization spreads out influence, while L1 encourages sparsity, which eliminates weak features
- Experts are akin to a post-hoc L1/L0 regularization
- Adding L1 to training is not sufficient because SGD implicitly results in a form of L2 regularization

**Note that these claims are all intuitive rather than rigorous, so take them with a grain of salt*