Fighting Sampling Bias in ML Models in Credit Scoring

N Kozodoi, M Alamgir, Y Gatsoulis, S Lessmann, L Moreira-Matias, K Papakonstantinou

Presentation Outline

1. Sampling Bias Problem

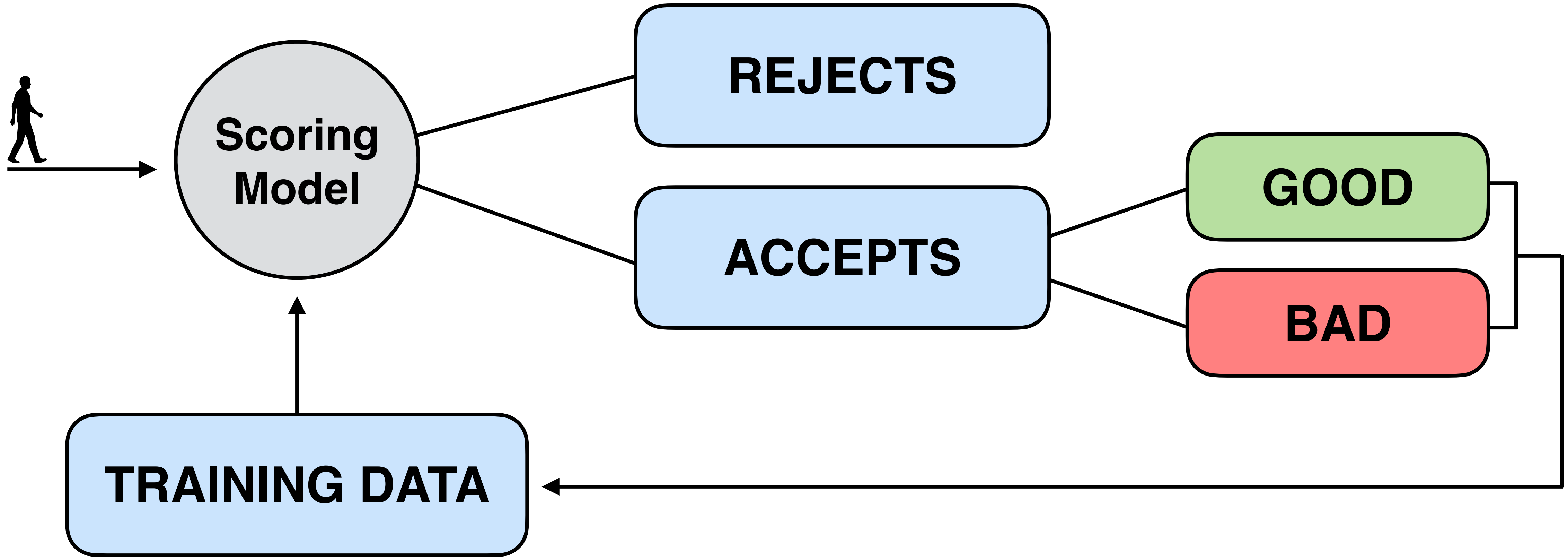
- Problem setup & illustration
- Impact on ML model training and evaluation

2. How to Correct Sampling Bias?

- Improving training under sampling bias
- Improving evaluation under sampling bias

3. Further Challenges

Acceptance Loop in Credit Scoring

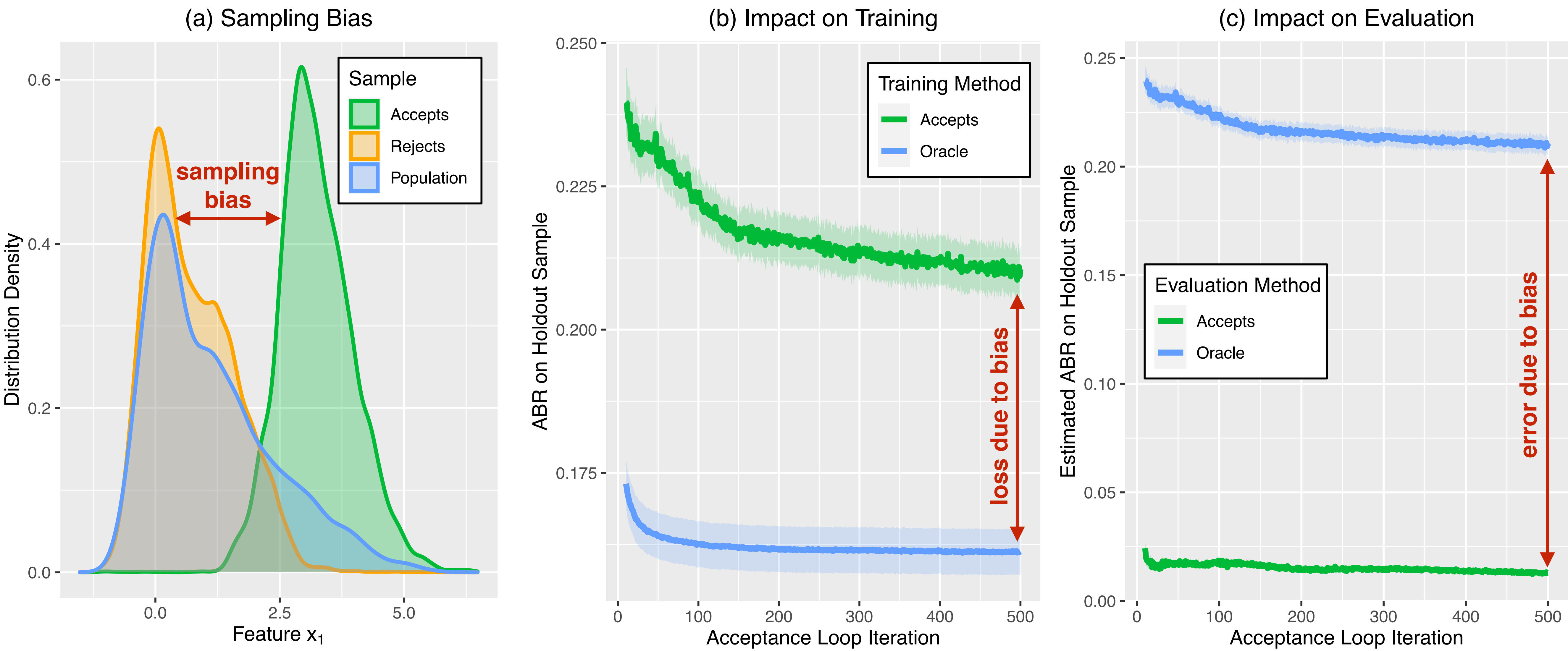


- **scoring model filters incoming loan applications**
  - ML model observes features of incoming applicants
  - predicts whether an applicant will repay the loan
- **training a model requires data with known outcomes**
  - outcomes are only observed for previously **accepted applicants**
  - labels are missing **not completely at random** but depending on the model
- **sampling bias may amplify with acceptance loop iterations**

Sampling Bias Illustration

Sampling bias in accepts affects model training and evaluation:

- training a model on a biased sample **decreases its performance**
- evaluating a model on a biased sample provides a **misleading estimate**



ABR = average **BAD** rate among accepts; lower is better

Presentation Outline

1. Sampling Bias Problem
   - Problem setup & illustration
   - Impact on model training and evaluation

2. **How to Correct Sampling Bias?**
   - Improving training under sampling bias
   - Improving evaluation under sampling bias

3. Further Challenges

Training under Sampling Bias

How to improve training?

Data augmentation (label rejects)	Extract information from rejects
• **label rejects** using a certain technique • **augment training data** of **accepts** with pseudo-labeled **rejects** • use augmented data for training • e.g., **label all rejects as BAD**	• estimate **distribution mismatch** between **accepts** and target population • account for the mismatch during training without explicitly labeling **rejects** • e.g., **reweighting the loss**

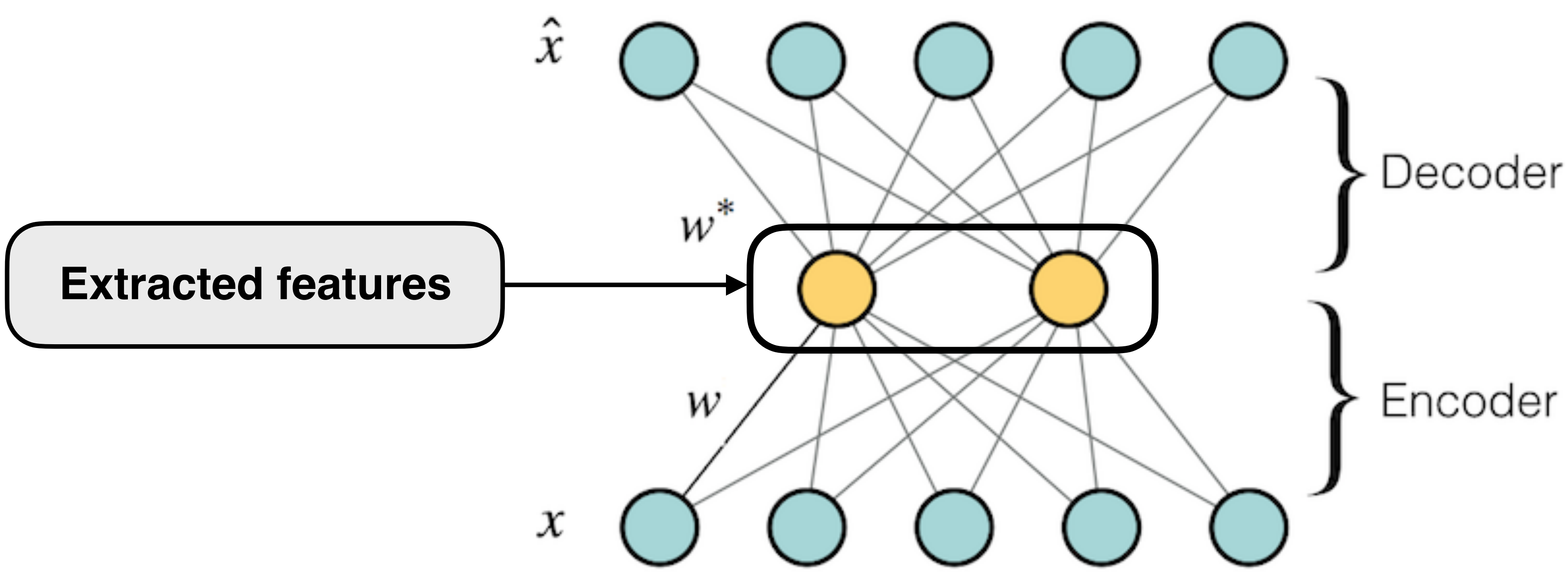
Extracting Information: Autoencoders

Idea:

- Use rejects to extract useful features **without labeling them**

Pipeline:

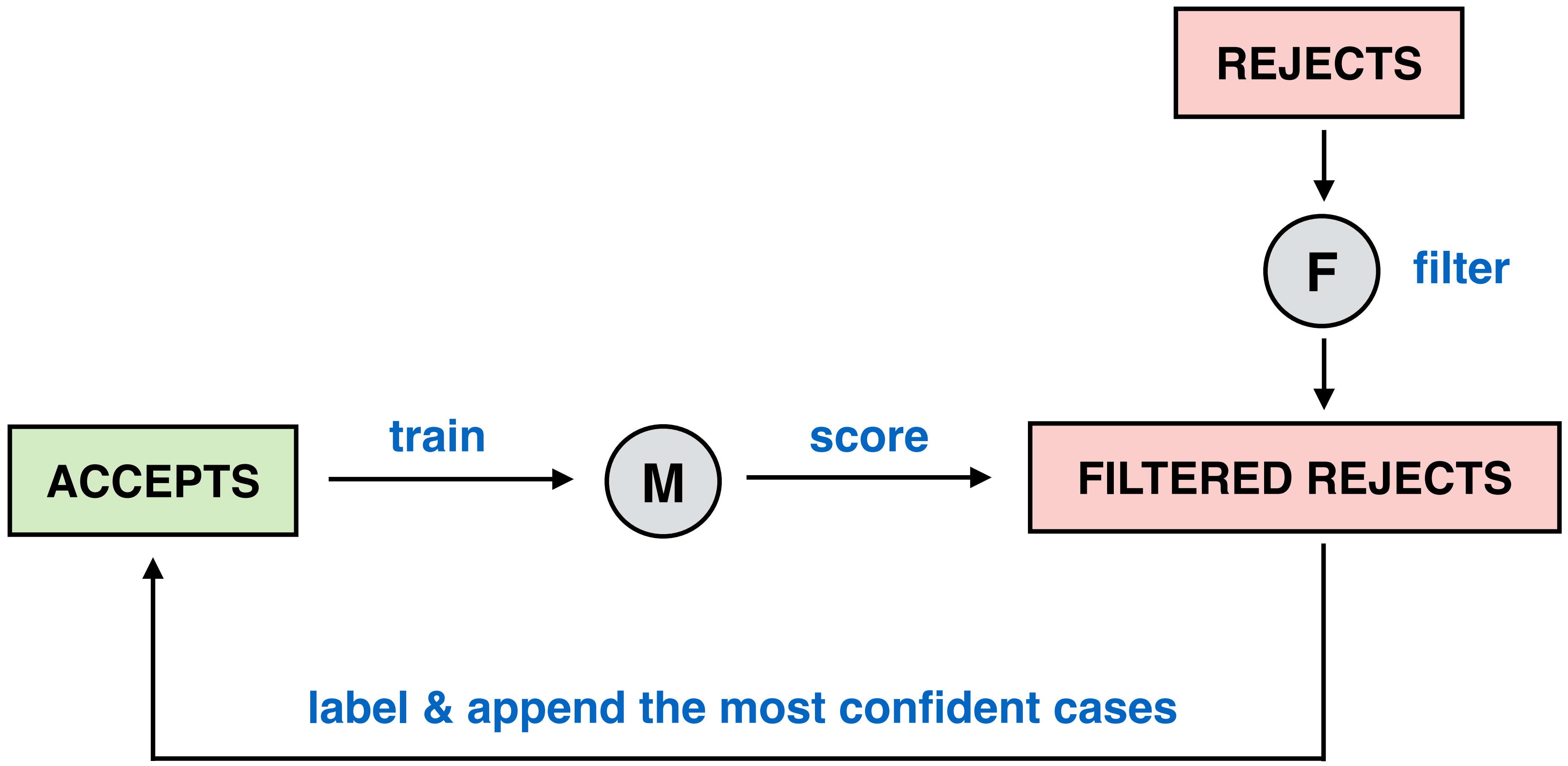
- Train Autoencoder on **accepts** + **rejects**
- Add distribution **mismatch penalty** to the loss function
- Use a bottleneck layer to **extract features**
- Append new features to accepts and train a new model



Labeling: Bias-Aware Self-Learning

Pipeline:

- iteratively label **selected rejects** using predictions from a weak classifier
- implement **multiple techniques** to reduce the risk of error propagation
  - filtering **rejects** coming from the most different distribution region
  - using imbalance multiplier to label & append more **BAD** applicants
  - early stopping labeling iterations to avoid overfitting on **accepts**



Evaluation under Sampling Bias

How to improve evaluation?

Collect unbiased sample

- evaluate on a **representative sample** to avoid sampling bias
- requires issuing loans to **random set of applicants** without scoring
- **issue:** very costly to set up

Adjust evaluation framework

- use techniques to account for the **distribution mismatch**
- incorporate **rejects** into evaluation
- **issue:** labels of **rejects** are unknown

Bayesian Evaluation Framework

- estimating evaluation metric ***M*** on a set ***S*** containing:
  - **accepts** with the true labels
  - **rejects** with random pseudo-labels based on the prior P(**BAD**)
- estimate prior P(**BAD**) based on the **current scorecard** $f(X)$

input : model $f(X)$, evaluation sample S consisting of labeled accepts $S^a = \{(\mathbf{X}^a, \mathbf{y}^a)\}$ and unlabeled rejects $\mathbf{X}^r$, prior $\mathbf{P}(\mathbf{y}^r | X^r)$, evaluation metric $M(f, S, \tau)$, meta-parameters j_{max}, ϵ

output: Bayesian evaluation metric $\mathrm{BM}(f, S, \tau)$

$j = 0$; $\Delta = \epsilon$; $E^c = \{\}$; `// initialization`

while $(j \leq j_{max})$ *and* $(\Delta \geq \epsilon)$ **do**

    $j = j + 1$

    $\mathbf{y}^r = \mathrm{binomial}(1, \mathbf{P}(\mathbf{y}^r | \mathbf{X}^r))$; `// generate labels of rejects`

    $S_j = \{(\mathbf{X}^a, \mathbf{y}^a)\} \cup \{(\mathbf{X}^r, \mathbf{y}^r)\}$; `// construct evaluation sample`

    $E^c_j = \sum_{i=1}^{j} M(f(X), S_i, \tau)/j$; `// evaluate`

    $\Delta = E^c_j - E^c_{j-1}$; `// check convergence`

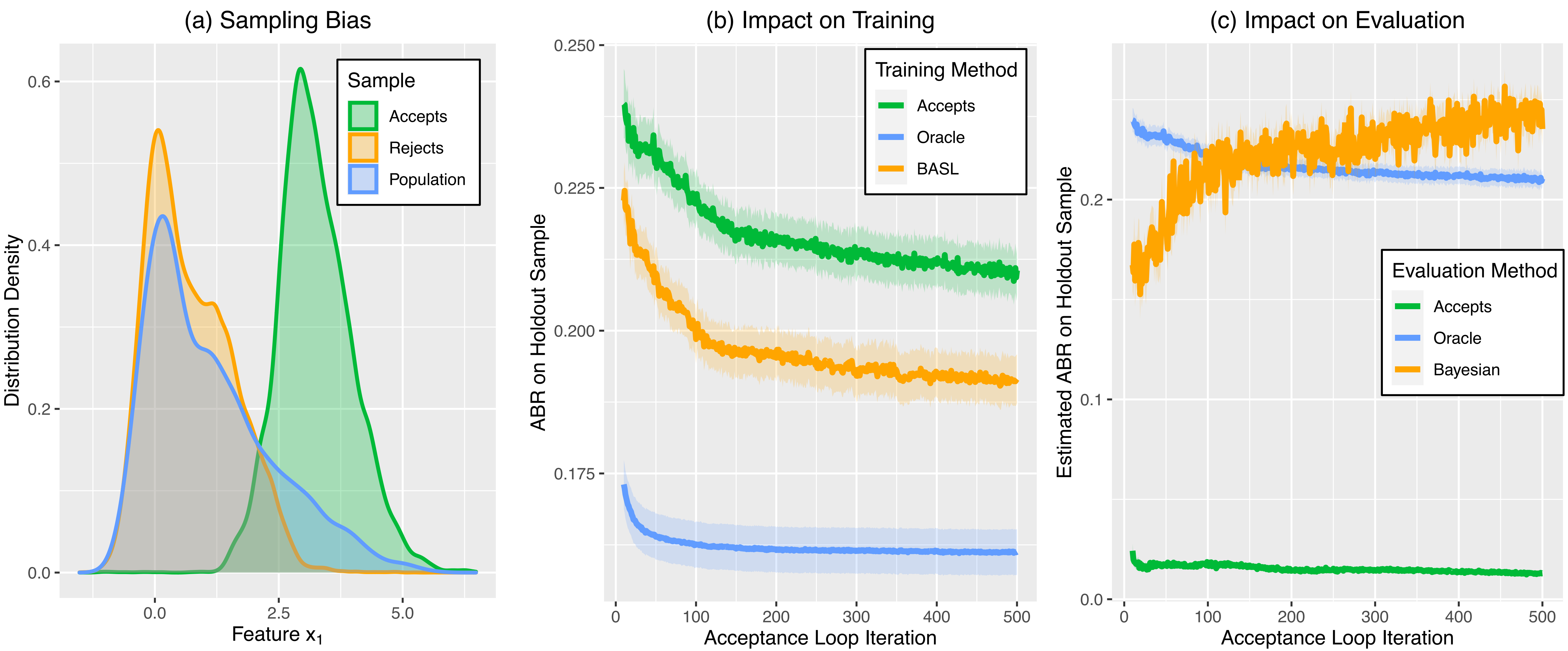
end

return $BM(f, S, \tau) = E^c_j$

Potential Performance Gains

Using bias correction methods allows to partly recover loss due bias

- **improving performance** of the model on new applications
- **improving performance estimate** of the model on new applications



Presentation Outline

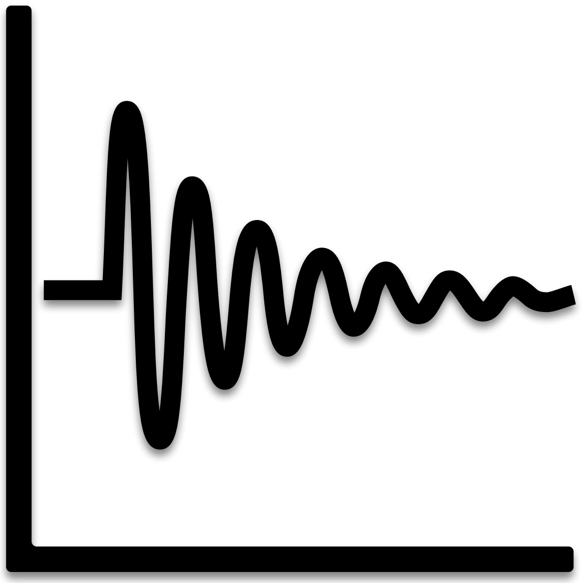
1. **Sampling Bias Problem**
   - Problem setup & illustration
   - Impact on model training and evaluation

2. **How to Correct Sampling Bias?**
   - Improving training under sampling bias
   - Improving evaluation under sampling bias

3. **Further Challenges**

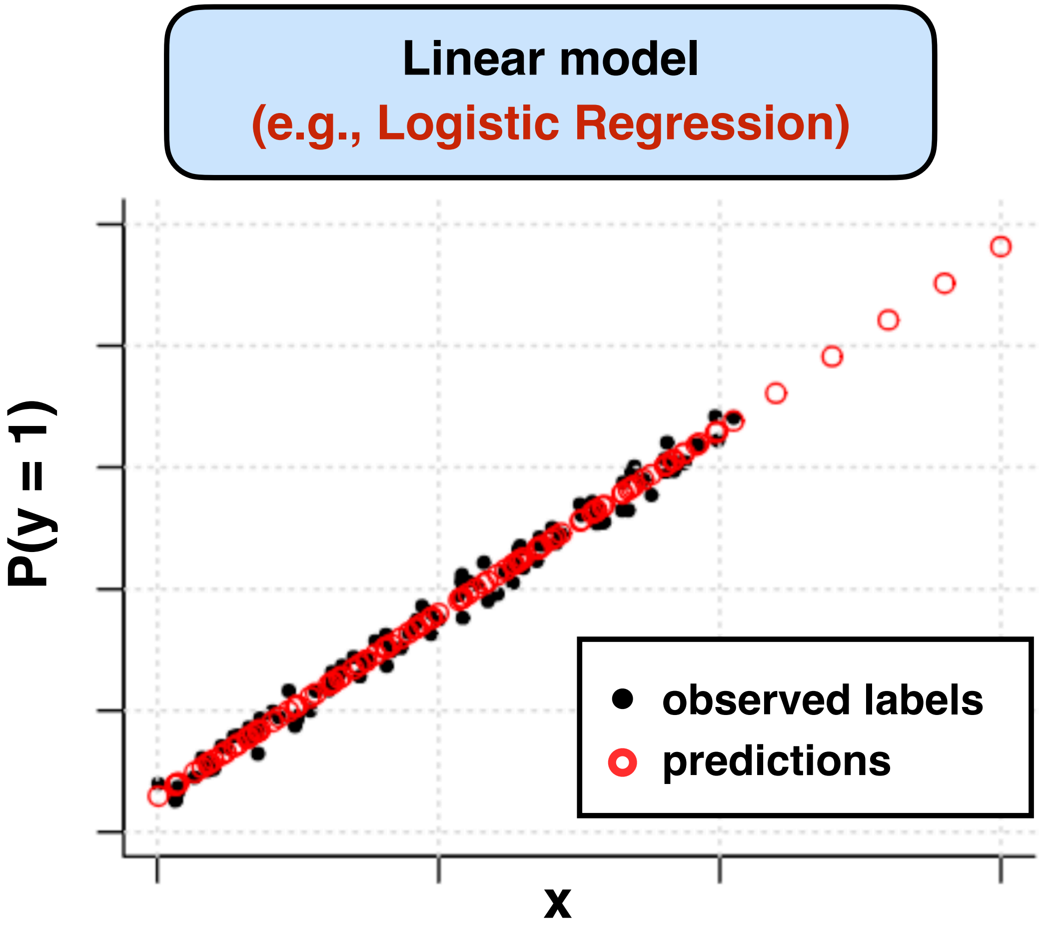
Dataset Shift and Sampling Bias

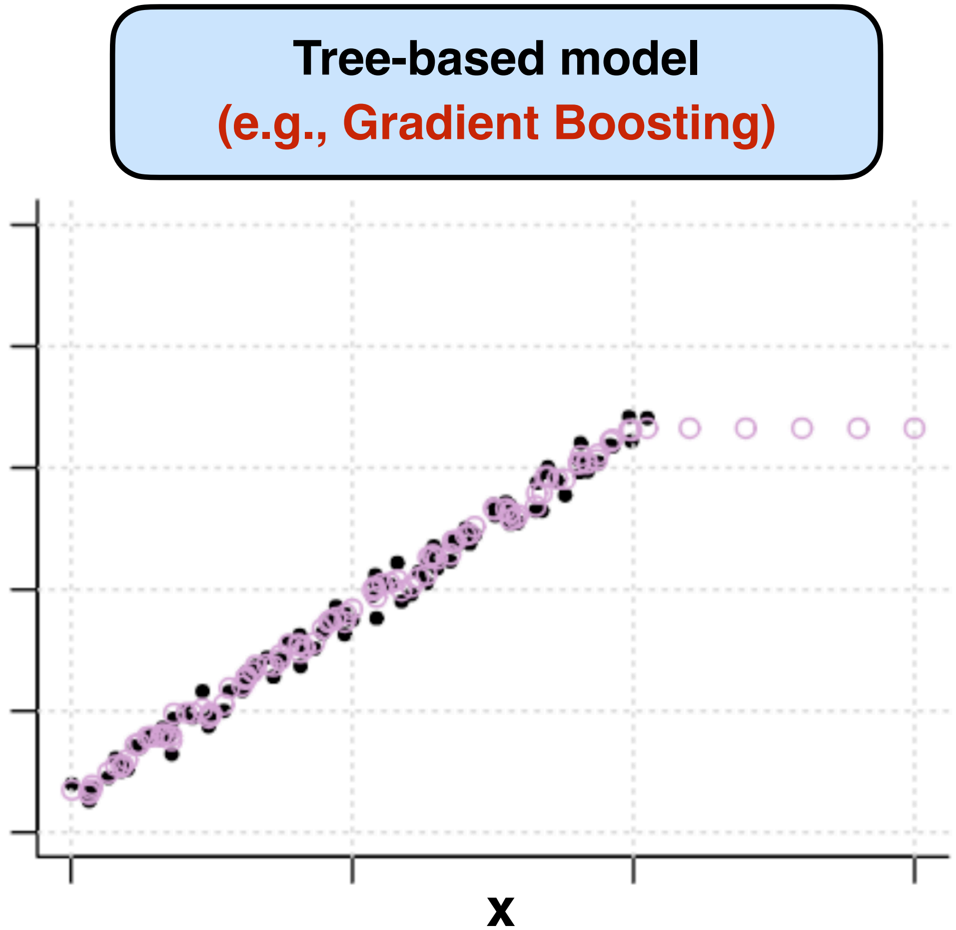
- **distribution discrepancy is also affected by dataset shift**
  - complicates the correction of sampling bias between **accepts**/**rejects**
  - long delay between accepting an applicant and learning their label
- **covariate shift**
  - change in the feature distribution between train and test data
  - e.g., changes in the acceptance policy or marketing strategy
- **concept shift**
  - change in the functional feature-target relationship
  - e.g., changes in the business cycle



Sampling Bias in Different Environments

- **magnitude of sampling bias depends on many factors**
- **lower approval rates => stronger bias**
  - low acceptance increases difference between **accepts** and population
  - can make it too difficult for bias correction to work given a sparse sample
- **classifiers have different extrapolation abilities**

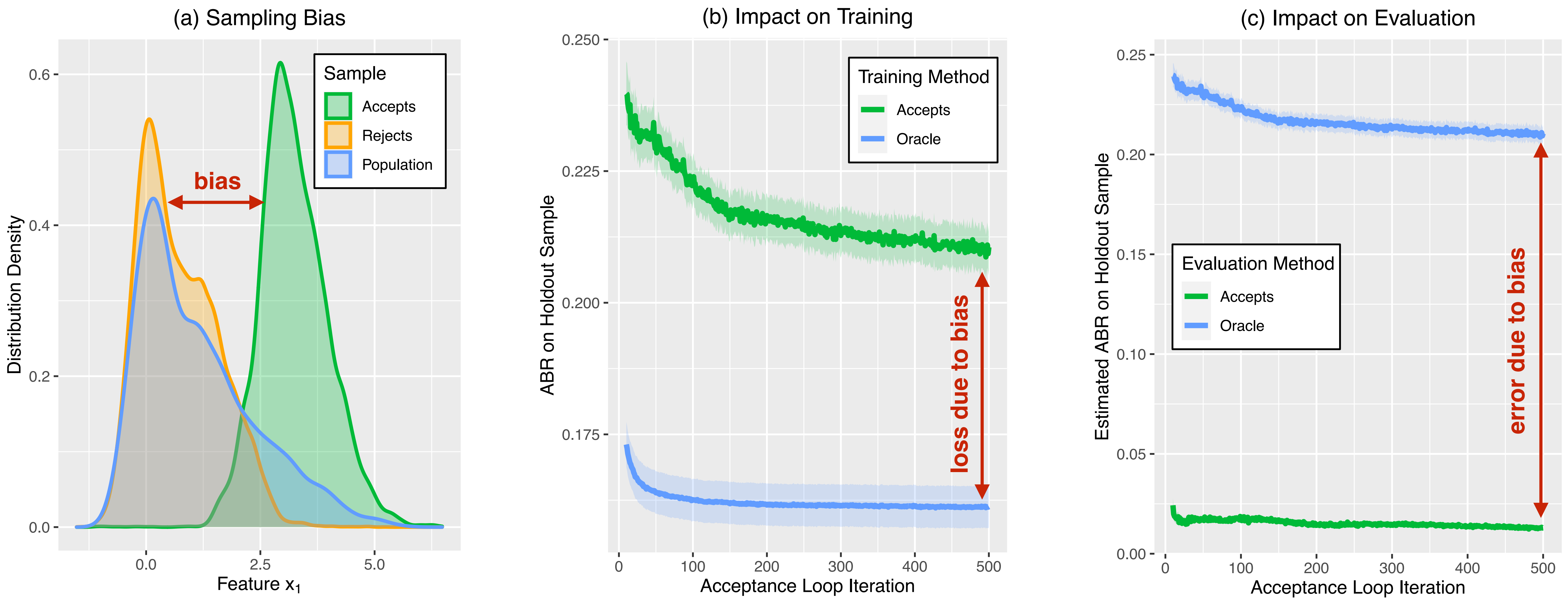




Some Further Challenges

- **regulation-related challenges**
  - keeping data on **rejected applicants** might not be feasible
  - need to create synthetic samples similar to real **rejects**
- **bias illustration in ML models**
  - detecting bias in non-parametric models is not straightforward
  - need to illustrate bias through the lens of performance / model predictions

Thanks for your Attention!



Contact:

n.kozodoi@icloud.com

www.linkedin.com/in/kozodoi

www.kozodoi.me

Slides: