# Optimal Transport

## Learning with a Wasserstein Loss

Nicolás Violante

# Introduction

# Ordinal classification

- Particular case of multi-class classification where the labels have a natural ordering, i.e. ground truth metric

- Example: classify movies as bad (0), good (1), excellent (2)

# Wasserstein distance as loss function

- Optimal transport naturally lifts a ground truth metric to a distance between probability vectors
- Use Sinkhorn algorithm to compute the loss between network outputs and targets
- Approximate the gradients via the dual problem and backpropagate

# Dataset

# Car Dataset

- 1728 examples of cars labeled as unacceptable (1210), acceptable (384), good (134)

- Each example has 6 attributes

- Balanced dataset with 134 examples of each class

# Experiments

# Experiments

- Sinkhorn iterations
- Cost matrix
- Small network
- Unbalanced dataset
- Sinkhorn divergence

# Sinkhorn iterations

- 350 slower than Cross Entropy for N=10

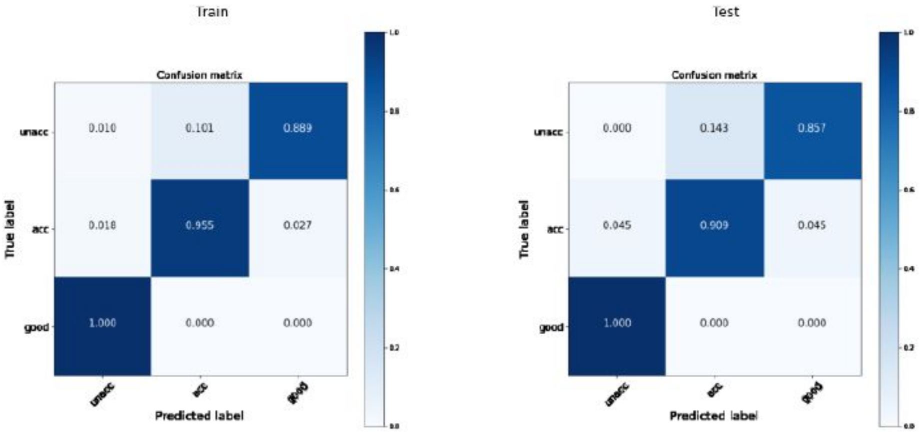
| Sinkhorn Iterations | Accuracy | Average loss time computation (sec) |
|---|---|---|
| 1 | 0.963 | 0.0446 |
| 10 | 0.988 | 0.0448 |
| 100 | 0.963 | 0.0545 |
| 500 | 0.988 | 0.089 |
| 1000 | 0.963 | 0.153 |
| 2000 | 0.988 | 0.223 |

# Cost matrix

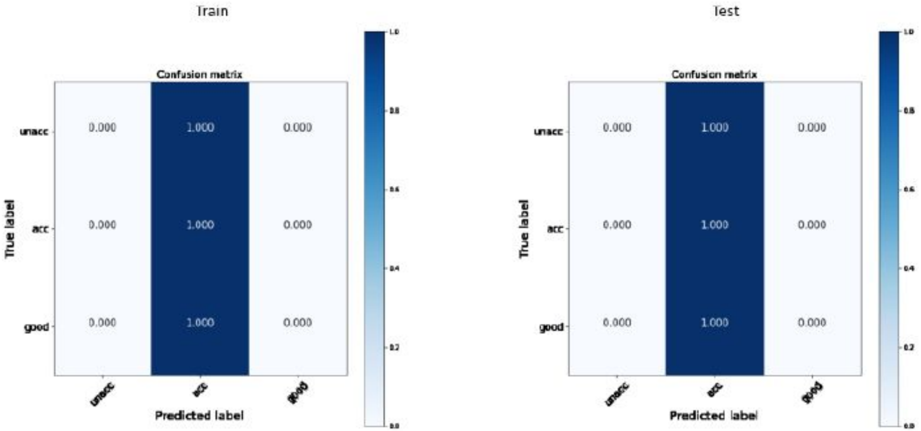
- We can control the weight of each miss-classification error via the cost matrix. For example (toy example)

$$C_1 = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

# Cost matrix



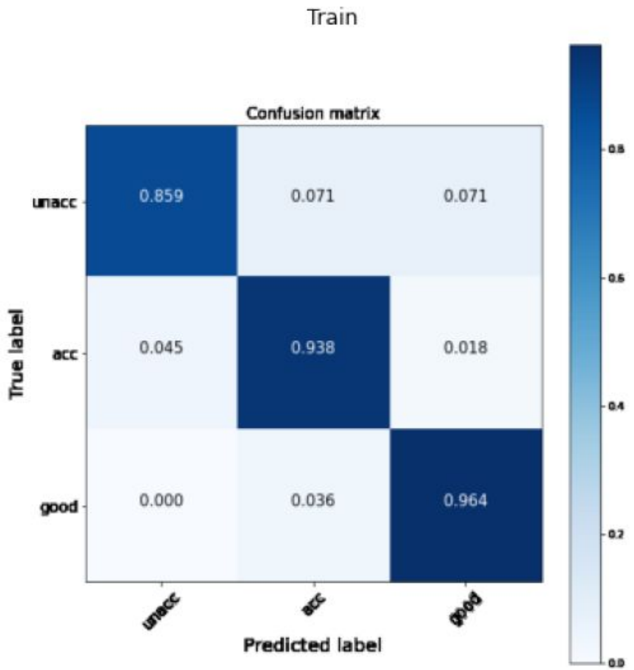

(a) Results for cost $C_1$

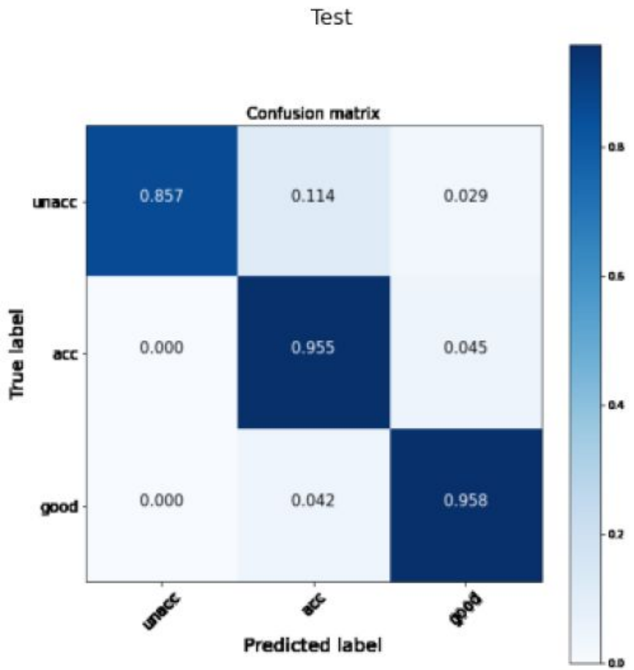



(b) Results for cost $C_2$

# Small network

- A single linear layer between inputs and outputs

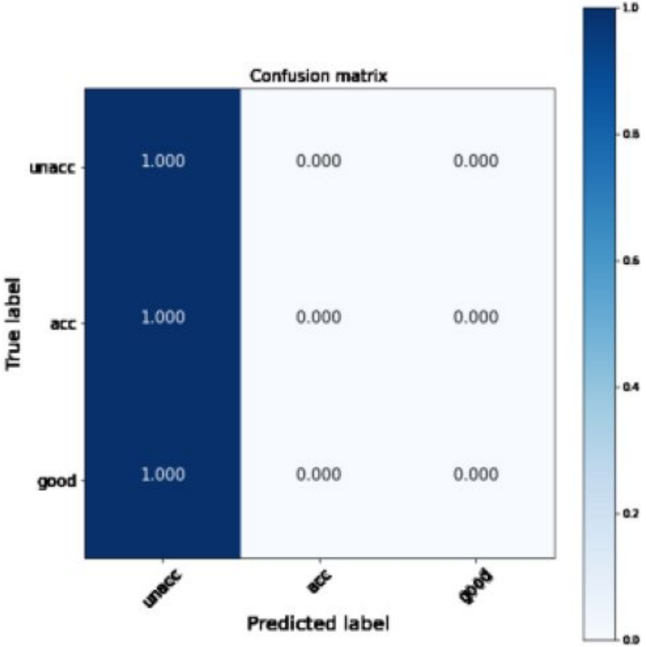


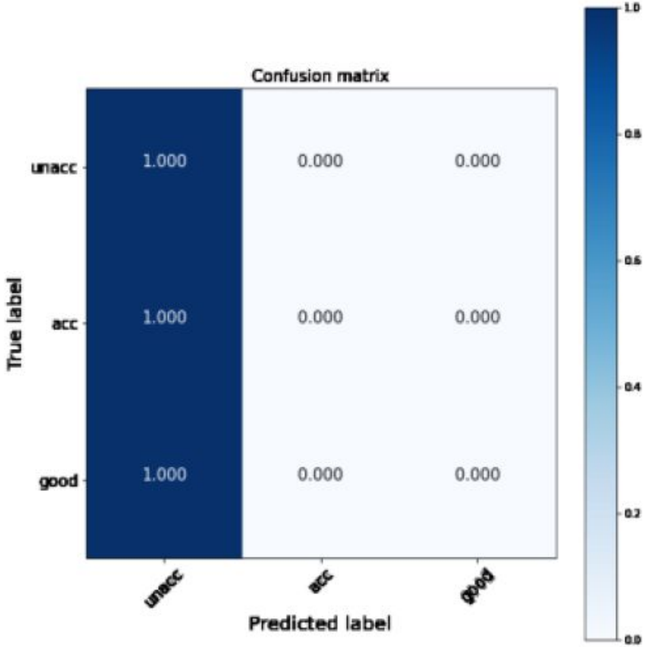
# Unbalanced dataset

- The predictions collapse into the dominant class

Train



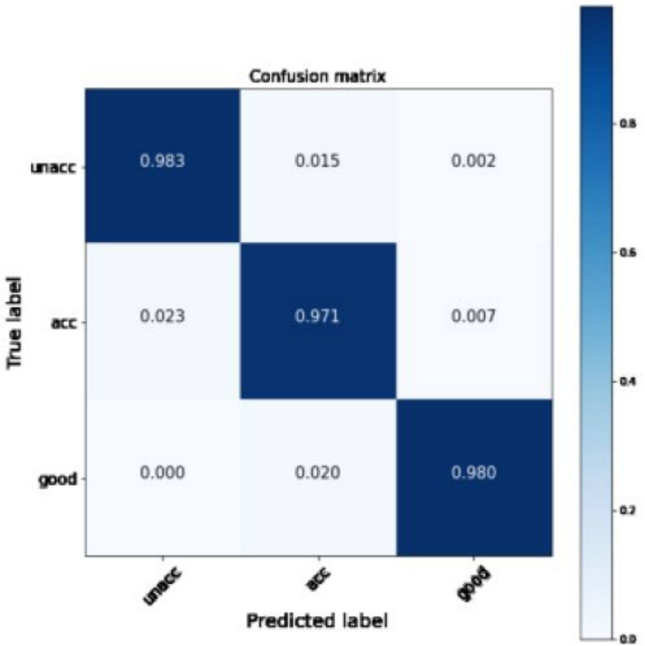

Test



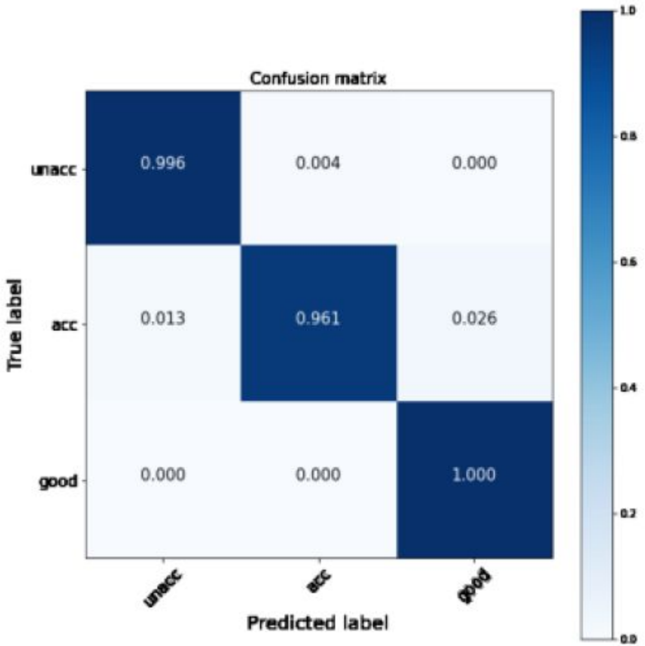
# Sinkhorn divergence

- Helps to prevent collapse (most of the time)

Train




Test



# Conclusions

# Conclusions

- Wasserstein distance can be used to train ordinal classification models and obtain competitive results
- The cost matrix C offers a natural way to penalize different errors with different weights
- Additional computational cost for running Sinkhorn may be prohibitive for large scale training
- Collapse with unbalanced dataset

Thank you