Minimax Generalized Cross-Entropy

Kartheek Bondugula[1] Santiago Mazuelas[1,2]
Aritz Pérez[1] Anqi Liu[3]

Classification performance depends on the loss function

Goal: Efficiently learning of accurate, robust and calibrated classifiers

- Mean absolute error (**MAE**):
  - **robust to label noise**
  - **Non-convex objective**, **poor probability calibration**
- Cross-entropy (**CE**):
  - **Sensitive to label noise**
  - **Convex objective**, **well-calibrated probabilities**

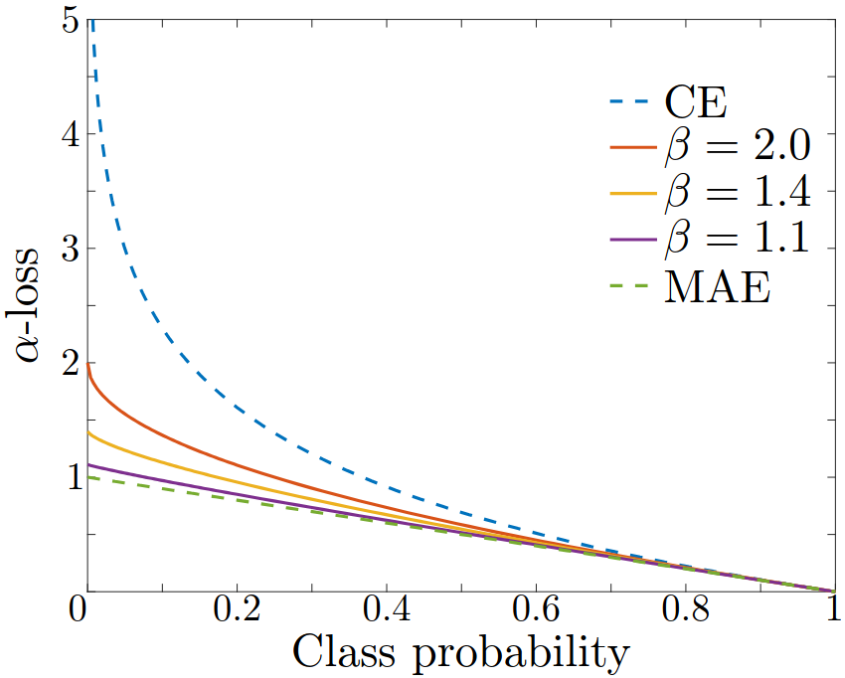
Question: Can we balance accuracy, calibration, and tractability?

A tunable loss function

α-loss: **interpolates** between CE ($\beta = 1$) and MAE ($\beta = \infty$)

$$\ell_\beta(h, (x, y)) = \beta \left(1 - h(y|x)^{1/\beta}\right)$$

h : classifier
(x, y) : labeled instance
$\beta = \frac{\alpha}{\alpha - 1}$: reparametrization , $\beta \in [1, \infty)$
$h(y|x)$: probability of class y given x



β controls the trade-off between **robustness** and **optimization**

Generalized Cross-Entropy (GCE)

- Empirical risk minimization with α-loss
- **Non-convex** objective

Our contribution

Minimax GCE: **Minimax approach** using **α-loss**

$$\min_{h} \max_{p \in \mathcal{U}} \mathbb{E}_{\mathrm{p}} \ell_{\beta}(h, p)$$

Minimize the risk against the **worst-case distribution** in the **uncertainty set** $\mathcal{U}$

Key advantages:

- **Convex** optimization for $\beta \in [1, \infty)$
- Efficient learning using **stochastic gradient descent**
- Learn **accurate, calibrated and robust** classifiers

Convex optimization problem

Uncertainty set

$$\mathcal{U} = \{\mathrm{p} : \ |\mathbb{E}_{\mathrm{p}}\{\Phi(x, y)\} - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda}, \ \mathrm{p}(x) = \mathrm{p}^*(x)\}$$

Φ: feature mapping $\quad \boldsymbol{\tau}$: estimate $\quad \boldsymbol{\lambda}$: confidence $\quad \mathrm{p}^*(x)$: true marginal

MCGE convex optimization problem

$$\min_{\boldsymbol{\mu} \in \mathbb{R}^m} \quad -\boldsymbol{\tau}^\top \boldsymbol{\mu} + \boldsymbol{\lambda}^\top |\boldsymbol{\mu}| - \mathbb{E}_{\mathrm{p}^*(x)}\left[\varphi_\beta(x, \boldsymbol{\mu})\right]$$

$$\text{s.t.} \ \sum_y \left(\frac{f(x, \boldsymbol{\mu})_y + \varphi_\beta(x, \boldsymbol{\mu})}{\beta} + 1 \right)_+^\beta = 1$$

- $\mathcal{U}$ induces **regularization** and a margin-based classifier
- φ_β is **implicitly defined**, and **links** the **classifier** with the **worst-case distribution**

Efficient optimization

Stochastic Gradient Descent

The gradient of the MGCE objective at (x, y)

$$\boldsymbol{\lambda} \odot \operatorname{sign}(\boldsymbol{\mu}) - \Phi(x, y) - \frac{\partial \varphi_\beta(x, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$$

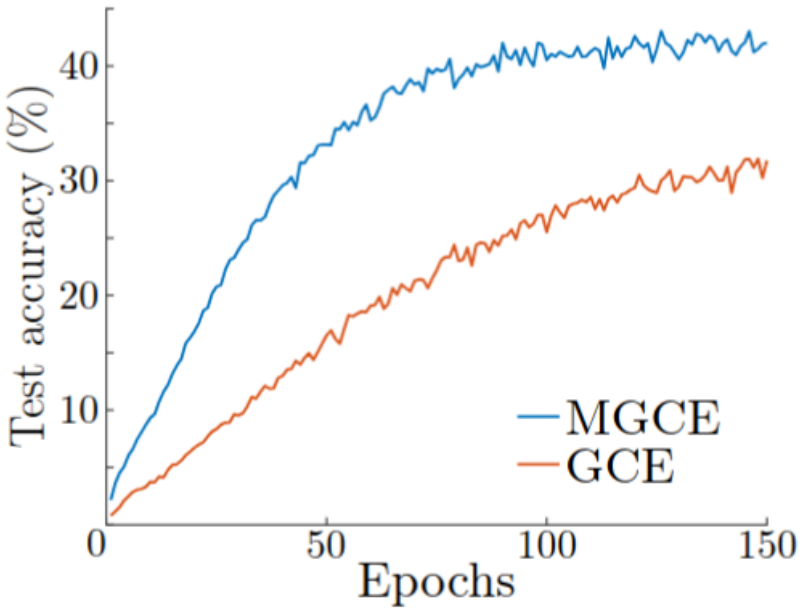
Implicit differentiation of φ_β

$$\frac{\partial \varphi_\beta(x, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -\sum_{y=1}^{k} \mathrm{p}_\beta(y|x; \boldsymbol{\mu})\, \Phi(x, y)$$

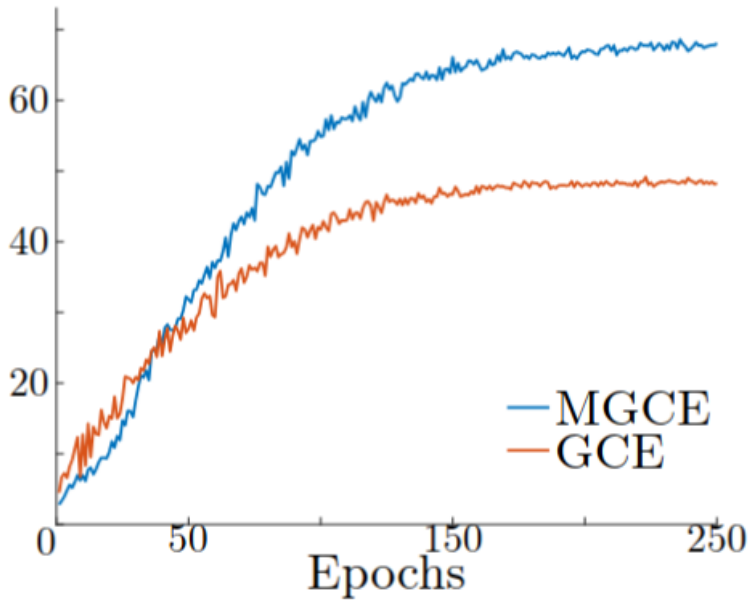
$\mathrm{p}_\beta(y|x; \boldsymbol{\mu})$: worst-case distribution for $h(y|x; \boldsymbol{\mu})$

Efficient computation via bisection

Fast convergence, $\beta = 1.4$



Tiny ImageNet



WebVision (label noise)

Accurate, calibrated and robust classifiers, $\beta = 1.4$

Dataset	Method	Accuracy (%)		SCE (%)	
		η=0.0	η=0.4	η=0.0	η=0.4
FashionMNIST	MGCE	**92.94**	88.65	**5.20**	**2.54**
	GCE	92.93	**89.74**	5.52	8.63
	CE	92.73	87.40	5.33	33.52
	MAE	91.84	88.44	73.47	74.23
CIFAR10	MGCE	**91.01**	83.49	**5.53**	**12.15**
	GCE	90.55	**83.80**	6.14	12.95
	CE	90.80	77.31	5.71	25.31
	MAE	89.31	82.64	73.00	68.29
SVHN	MGCE	**95.04**	**92.92**	3.19	**3.96**
	GCE	94.68	92.84	**2.58**	5.27
	CE	93.49	90.88	4.22	32.61
	MAE	92.89	92.32	75.76	77.91
CIFAR100	MGCE	64.35	**52.65**	**22.22**	23.54
	GCE	**64.82**	51.37	22.89	37.16
	CE	64.66	42.32	22.19	**6.06**
	MAE	62.17	41.99	60.47	60.46

SCE: static calibration error

Thank you!

Minimax Generalized Cross-Entropy

For more details, please visit
Poster #185 today at **15:00**