PENEX: AdaBoost-Inspired Neural Network Regularization

Friday Talk

Klaus-Rudolf Kladny, Bernhard Schölkopf, Michael Muehlebach

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Tübingen AI Center

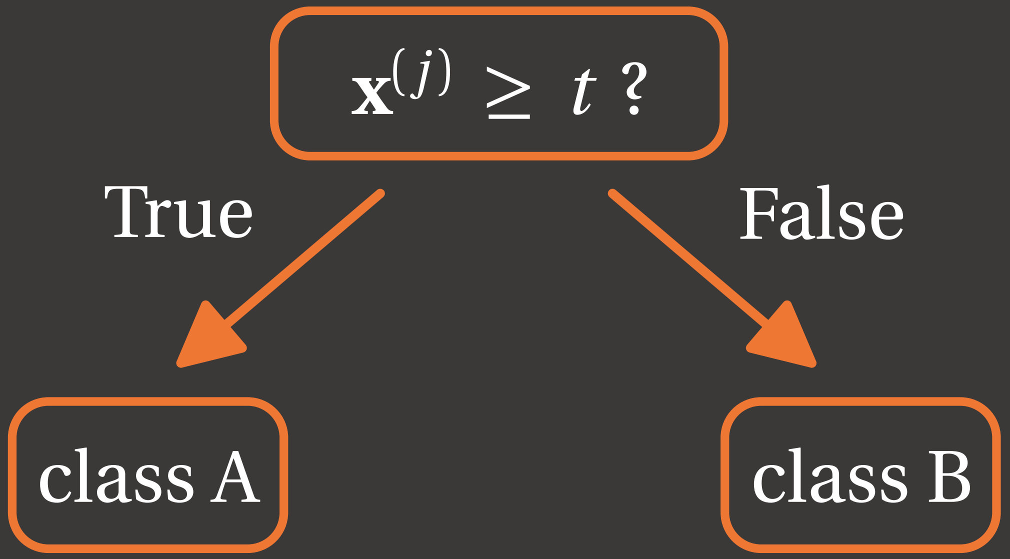
ellis | INSTITUTE TÜBINGEN

What loss do you use to train your classifier?*

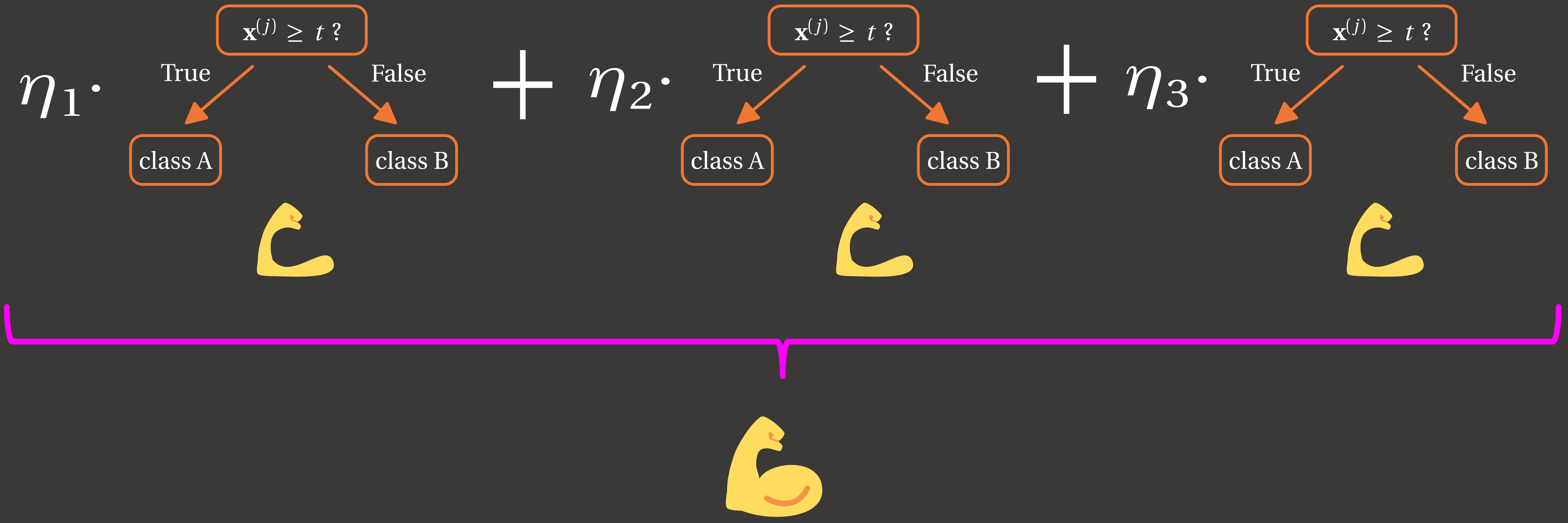
*Yes, an LLM qualifies as a classifier

What is AdaBoost? 🚀

Weak learner



Constructing a strong learner



Optimization objective: exponential loss

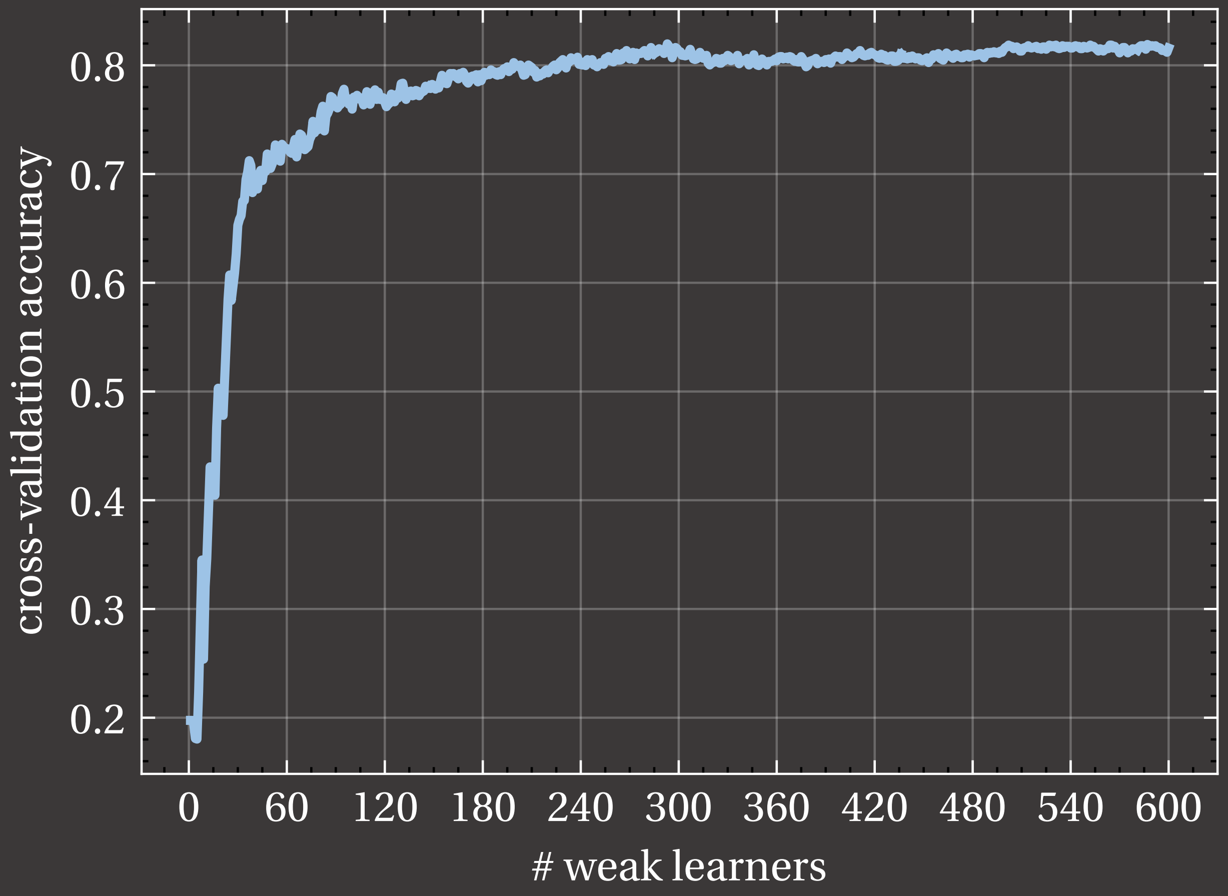
$$f(\mathbf{x}) = \sum_{i=1}^{N} \eta_i g_i(\mathbf{x})$$

g_i greedily minimize

$$\hat{\mathbb{E}}\left[\exp\left\{-y f(\mathbf{x})\right\}\right], \qquad y \in \{-1, 1\}$$

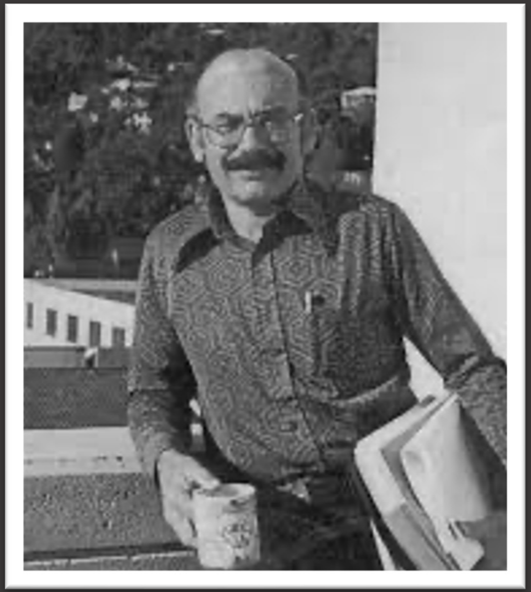
Breiman, Leo. "Prediction Games and Arcing Algorithms." *Neural Computation* 11.7 (1999): 1493-1517.

AdaBoost is resilient to "overfitting"



“[AdaBoost with trees is] the best off-the-shelf classifier in the world.”

- Leo Breiman



Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Additive logistic regression: a statistical view of boosting. *The Annals of Statistics* 28.2 (2000): 337-407.

Can we translate the "AdaBoost magic" to neural networks?

Multiclass exponential loss

$$\hat{\mathbb{E}}\left[\exp\left\{-(K-1)^{-1} f^{(y)}(\mathbf{x})\right\}\right] \quad \text{subject to} \quad \sum_{j=1}^{K} f^{(j)}(\mathbf{x}) = 0,\ \forall \mathbf{x}.$$

exponential loss

prevents logits from diverging

Constraints 😟

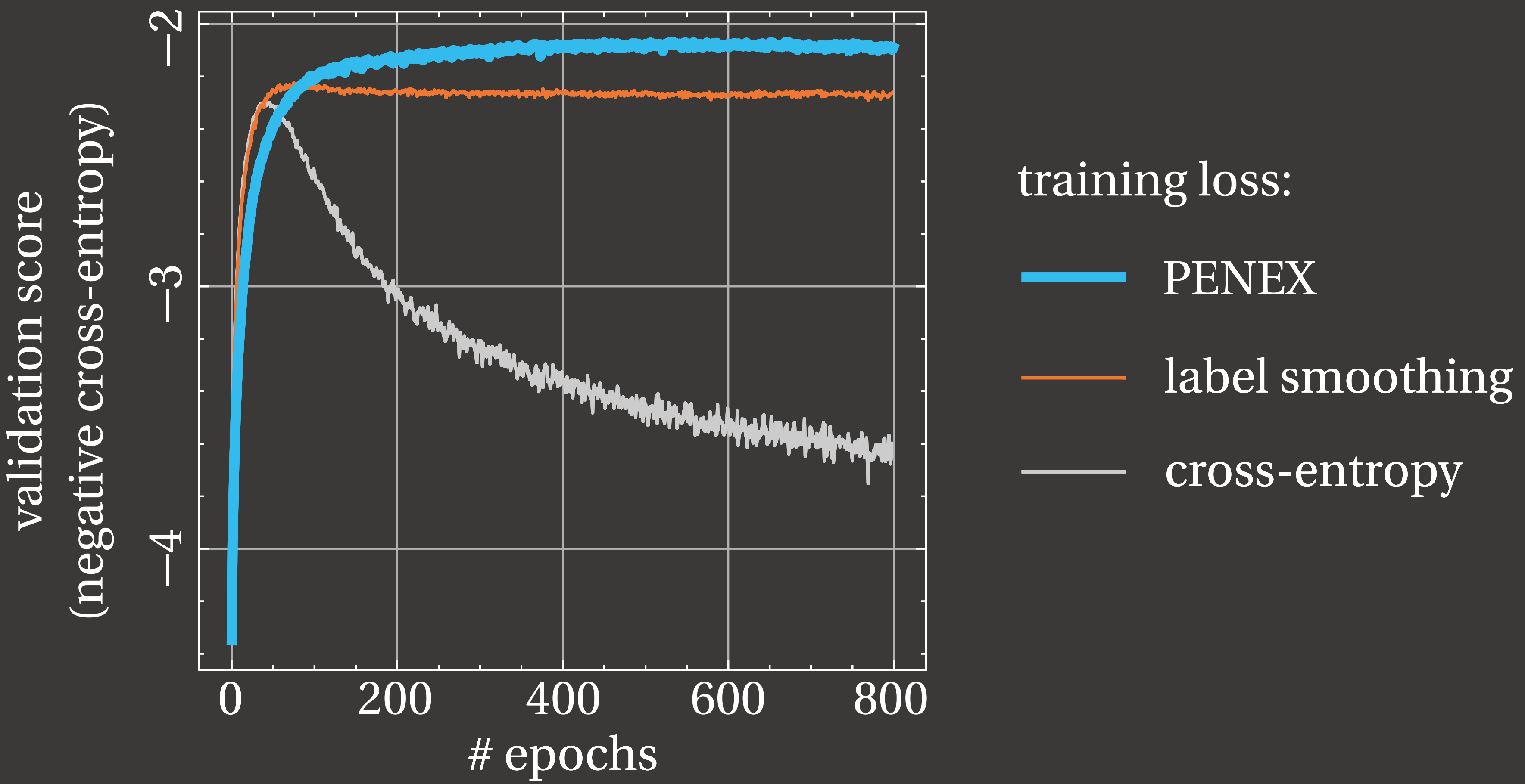
Zhu, Ji, et al. "Multi-class adaboost." *Statistics and its Interface* 2.3 (2009): 349-360.

Penalized exponential loss (PENEX)

$$\hat{\mathbb{E}}\Big[\underbrace{\exp\{-\alpha f^{(y)}(\mathbf{x})\}}_{\text{exponential loss}} + \underbrace{\rho\sum_{j=1}^{K}\exp\{f^{(j)}(\mathbf{x})\}}_{\text{prevents logits from diverging}}\Big]$$

No constraints! 🥳

... and it works 🔥



(CIFAR-100)

PENEX often works better than other methods

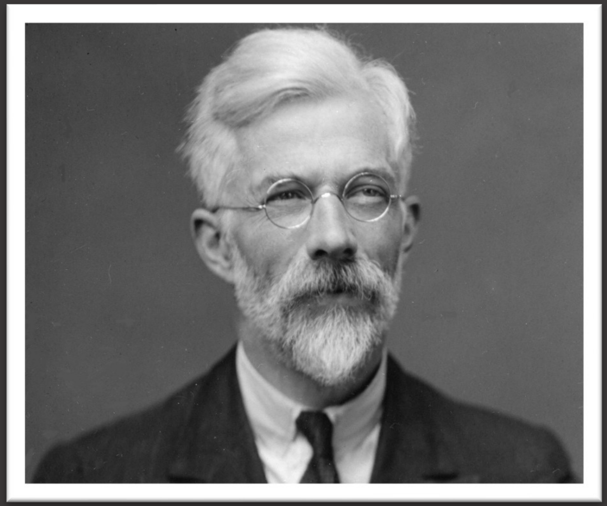
Method	Metric	CIFAR-10	Noisy CIFAR-10	CIFAR-100	PathMNIST	BBC News
CE	ACC	0.785 ± 0.004	0.724 ± 0.004	0.443 ± 0.004	0.826 ± 0.004	0.967 ± 0.007
	-ECE	-0.162 ± 0.003	-0.179 ± 0.003	-0.287 ± 0.003	-0.151 ± 0.004	$\mathbf{-0.032} \pm 0.006$
	-CE	-1.004 ± 0.024	-1.125 ± 0.019	-3.072 ± 0.034	-2.018 ± 0.130	-0.109 ± 0.024
	-BRIER	-0.346 ± 0.006	-0.424 ± 0.006	-0.794 ± 0.006	-0.300 ± 0.007	-0.051 ± 0.011
label smoothing	ACC	0.789 ± 0.004	0.747 ± 0.004	0.451 ± 0.005	0.829 ± 0.004	0.970 ± 0.006
	-ECE	-0.112 ± 0.002	-0.183 ± 0.003	$\mathbf{-0.147} \pm 0.002$	-0.109 ± 0.002	-0.033 ± 0.006
	-CE	-0.657 ± 0.011	-0.889 ± 0.008	-2.292 ± 0.019	$\mathbf{-0.589} \pm 0.012$	-0.115 ± 0.022
	-BRIER	-0.300 ± 0.005	-0.384 ± 0.004	-0.692 ± 0.004	-0.255 ± 0.005	-0.049 ± 0.010
confidence penalty	ACC	0.786 ± 0.004	0.733 ± 0.004	0.449 ± 0.006	0.828 ± 0.004	$\mathbf{0.974} \pm 0.006$
	-ECE	-0.130 ± 0.002	-0.149 ± 0.003	-0.152 ± 0.002	-0.110 ± 0.003	-0.050 ± 0.005
	-CE	-0.731 ± 0.015	-0.866 ± 0.009	-2.254 ± 0.018	-0.917 ± 0.047	-0.094 ± 0.015
	-BRIER	-0.317 ± 0.005	-0.385 ± 0.004	-0.695 ± 0.005	-0.262 ± 0.005	$\mathbf{-0.042} \pm 0.008$
focal loss	ACC	0.778 ± 0.004	0.708 ± 0.004	0.428 ± 0.005	0.803 ± 0.004	0.970 ± 0.006
	-ECE	-0.117 ± 0.002	-0.165 ± 0.003	-0.161 ± 0.003	-0.112 ± 0.003	-0.051 ± 0.005
	-CE	-0.661 ± 0.010	-0.905 ± 0.008	-2.341 ± 0.022	-0.939 ± 0.050	$\mathbf{-0.092} \pm 0.014$
	-BRIER	-0.313 ± 0.005	-0.423 ± 0.004	-0.723 ± 0.005	-0.291 ± 0.006	$\mathbf{-0.042} \pm 0.008$
PENEX	ACC	$\mathbf{0.793} \pm 0.004$	$\mathbf{0.766} \pm 0.004$	$\mathbf{0.460} \pm 0.005$	$\mathbf{0.833} \pm 0.004$	0.968 ± 0.006
	-ECE	$\mathbf{-0.109} \pm 0.002$	$\mathbf{-0.131} \pm 0.002$	$\mathbf{-0.147} \pm 0.003$	$\mathbf{-0.100} \pm 0.003$	-0.034 ± 0.006
	-CE	$\mathbf{-0.646} \pm 0.012$	$\mathbf{-0.716} \pm 0.009$	$\mathbf{-2.140} \pm 0.018$	-1.200 ± 0.089	-0.124 ± 0.025
	-BRIER	$\mathbf{-0.299} \pm 0.005$	$\mathbf{-0.332} \pm 0.004$	$\mathbf{-0.685} \pm 0.004$	$\mathbf{-0.251} \pm 0.006$	-0.055 ± 0.011

Theoretical properties of PENEX

Fisher consistency

"When applied to the whole population the derived statistic should be equal to the parameter."

- Ronald A. Fisher



Fisher, Ronald A. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.594-604 (1922): 309-368.

PENEX is Fisher consistent

$$\hat{\mathbb{E}}\Big[\exp\left\{-\alpha f^{(y)}(\mathbf{x})\right\} + \rho \sum_{j=1}^{K} \exp\left\{f^{(j)}(\mathbf{x})\right\}\Big]$$

PENEX is Fisher consistent

$$\mathbb{E}\Big[\exp\left\{-\alpha f^{(y)}(\mathbf{x})\right\} + \rho \sum_{j=1}^{K} \exp\left\{f^{(j)}(\mathbf{x})\right\}\Big]$$

Minimize w.r.t. $f \rightarrow f_*$

$$P(y\,|\,\mathbf{x}) \propto \exp\left\{(1+\alpha) f_*^{(y)}(\mathbf{x})\right\}, \quad \forall \mathbf{x}$$

Common regularizers fail Fisher consistency

$$\mathcal{L}_{\mathrm{CE}}(f) + \lambda\Omega(f)$$

Encompasses label smoothing, L2 regularization, confidence penalty, ...

Intuition: Regularization term $\Omega(f)$ pushes the solution off the Bayes-optimal predictor

$$\mathcal{L}_{\text{PENEX}}(f)$$

(nice car with five seats)

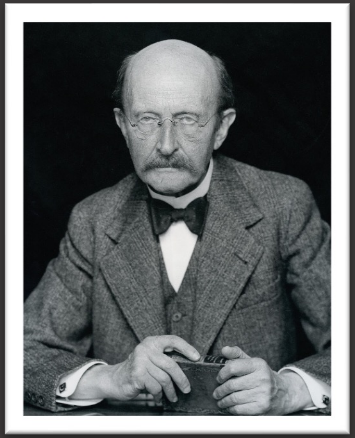
$$\mathcal{L}_{\text{CE}}(f) + \lambda\Omega(f)$$

(nice car with two seats and two extra seats mounted on top)

So why does it work?

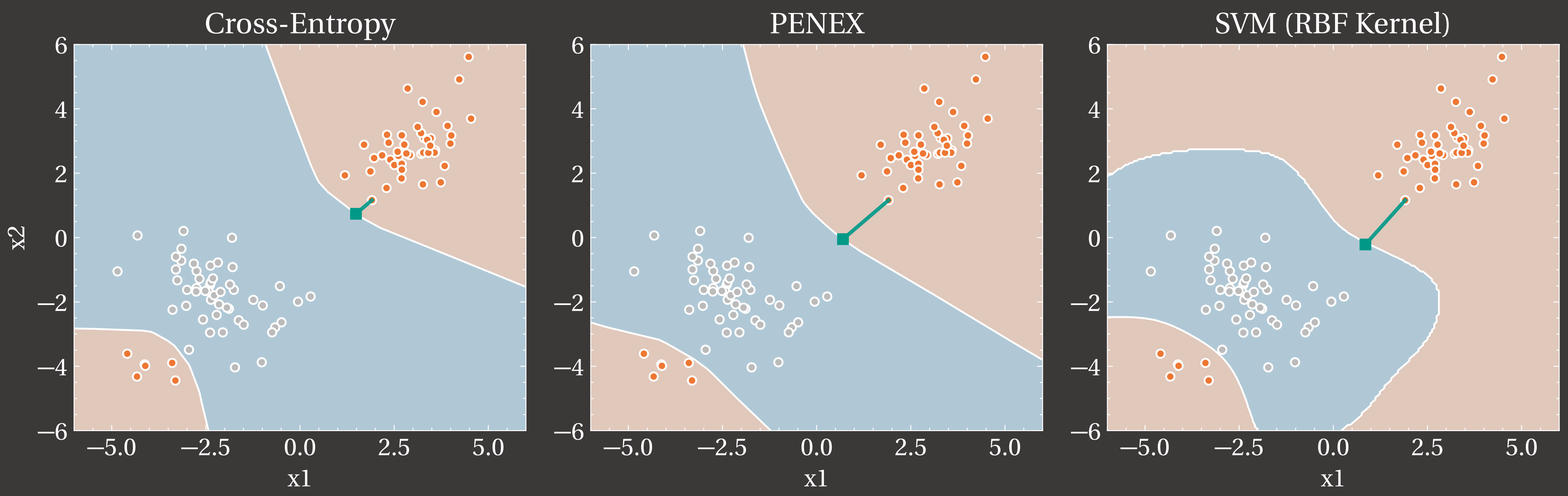
"Insight must precede application."

- Max Planck



https://www.mpi-cbg.de/research/our-research/basic-research

Implicit margin maximization



PENEX provably maximizes margins

Defining the margin for example $(\mathbf{x}, y)$ as

$$m_f(\mathbf{x}, y) := f^{(y)}(\mathbf{x}) - \max_{j \neq y} f^{(j)}(\mathbf{x}),$$

we show that

$$\mathbb{P}(m_f(\mathbf{x}, y) \leq \gamma) \leq e^{\gamma \frac{\alpha}{\alpha+1}} \rho^{-\frac{\alpha}{\alpha+1}} \mathbb{E}[\mathscr{L}_{\text{PENEX}}(f; \alpha, \rho)].$$

Extension of an earlier result: Schapire, Robert E., et al. "Boosting the margin: A new explanation for the effectiveness of voting methods." *The annals of statistics* 26.5 (1998): 1651-1686.

Key take-aways

- The "AdaBoost magic" can be translated to NNs
- Regularization is not at odds with Fisher consistency
- PENEX implicitly maximizes margins
- Let's question the very foundations!

Preprint:

Code:

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

E-Mail: kkladny@tuebingen.mpg.de