Uses and Abuses of the Cross-Entropy Loss: Case Studies in Modern Deep Learning

Elliott Gordon-Rodriguez, Gabriel Loaiza-Ganem, Geoff Pleiss, and John Cunningham

Columbia University; Layer6 AI

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

layer6

Cross-Entropy Loss

- The Cross-Entropy loss from predicting y with π is:

$$l(\pi; y) = -\sum_{k=1}^{K} y_k \log \pi_k \quad (1)$$

- It is equivalent to the (log) Categorical distribution:

$$p(y;\pi) = \prod_{k=1}^{K} \pi_k^{y_k} \quad (2)$$

- y lives in the discrete sample space of 1-hot vectors:

$$y \in \Omega_{\text{1-hot}} = \left\{ \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix}, \dots, \begin{pmatrix}0\\0\\\vdots\\1\end{pmatrix} \right\} \quad (3)$$

- What about $y \in \Delta^K$ (i.e. "soft labels")?

$$\Delta^K = \left\{ \pi \in \mathbb{R}_+^K : \sum_{k=1}^{K} \pi_k = 1 \right\} \quad (4)$$

The Continuous-Categorical

- $y \in \Delta^K$ follows a Continuous-Categorical (CC) distribution with parameter $\lambda \in \Delta^K$ if:

$$y \sim \mathcal{CC}(\lambda) \iff p(y;\lambda) = C(\lambda) \cdot \prod_{k=1}^{K} \lambda_k^{y_k} \quad (5)$$

- $C(\lambda)$ is just a normalizing constant:

$$C(\lambda) = \left((-1)^{K+1} \sum_{k=1}^{K} \frac{\lambda_k}{\prod_{i \neq k} \log \frac{\lambda_i}{\lambda_k}} \right)^{-1} \quad (6)$$

- The Continuous-Categorical log-likelihood defines a probabilistic loss function for $y \in \Delta^K$.
- Intuitively: a probabilistic version of the cross-entropy loss for simplex-valued data.
- Likelihood (5) can be optimized with auto-diff using the closed form expression (6).

Experiments: Label Smoothing

- Suppose we have a K-class classifier:

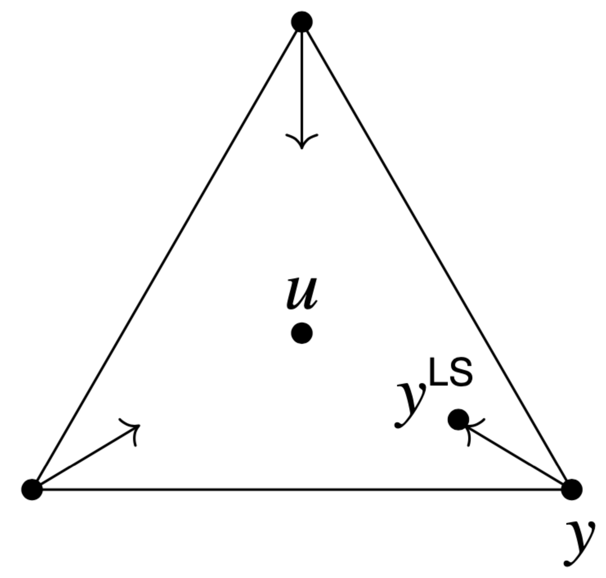
$$f_\theta : x \mapsto y$$

- Label Smoothing (LS):

$$y^{\text{LS}} = (1-\varepsilon)y + (\varepsilon/K)u$$

- Note the change in sample space:

$$y \in \Omega_{\text{1-hot}} \mapsto y^{\text{LS}} \in \Delta^K$$



Continuous-Categorical Label Smoothing (CC-LS)

- We replace the LS objective:

$$\min_\theta \mathcal{L}^{\text{LS}}(\theta) = -\sum_{(x,y)} \sum_k y_k^{\text{LS}} \cdot \log[f_\theta(x)]_k \quad (7)$$

- ... with the CC log-likelihood:

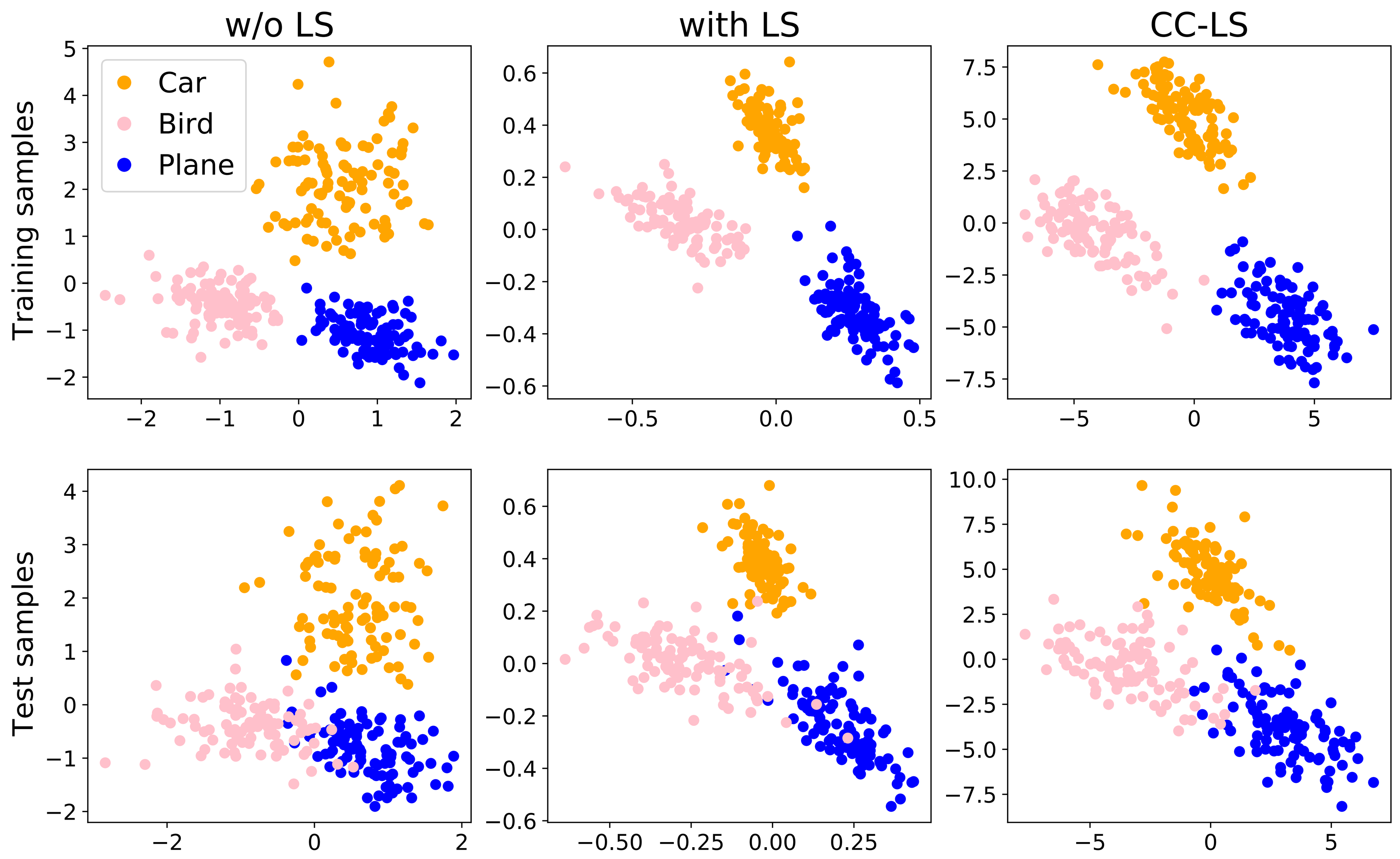
$$\min_\theta \mathcal{L}^{\text{CC-LS}}(\theta) = -\sum_{(x,y)} \left\{ \log C(f_\theta(x)) + \sum_k y_k^{\text{LS}} \cdot \log[f_\theta(x)]_k \right\}$$

- CC-LS regularizes more strongly than LS:

Table: Out-of-sample accuracy on CIFAR-10 (Alexnet)

Dropout	BatchNorm	w/o LS	with LS	CC-LS (ours)
No	No	86.8 (±0.1)	87.0 (±0.2)	**87.6** (±0.2)
Yes	No	87.0 (±0.2)	87.0 (±0.1)	**87.6** (±0.2)
No	Yes	**89.6** (±0.1)	89.2 (±0.1)	89.2 (±0.2)
Yes	Yes	**89.5** (±0.1)	89.1 (±0.2)	89.0 (±0.2)

- CC-LS learns richer hidden representations than LS:



Experiments: Actor-Mimic Reinforcement Learning

- Idea: train "Actor-Mimic Network" (AMN) to perform in multiple tasks by "imitating" pre-trained experts.
- For each state s, derive an "expert guidance" vector $y \in \Delta^K$
- Train $\pi_\theta^{\text{AM}} : s \mapsto a$, to optimize:

$$\min_\theta \mathcal{L}^{\text{AMN}}(\theta) = -\sum_{s,y} \sum_k y_k \cdot \log \pi_\theta(a_k|s) \quad (8)$$

- Again, $y \in \Delta^K$; targets are not one-hot but simplex-valued.

Continuous-Categorical Actor-Mimic Network (CC-AMN)

- We replace the AMN objective:

$$\min_\theta \mathcal{L}^{\text{AMN}}(\theta) = -\sum_{s,y} \sum_k y_k \cdot \log \pi_\theta(a_k|s) \quad (9)$$

- ... with the CC log-likelihood:

$$\min_\theta \mathcal{L}^{\text{CC-AMN}}(\theta) = -\sum_{s,y} \left\{ \log C(\pi_\theta(s)) + \sum_k y_k \cdot \log \pi_\theta(a_k|s) \right\} \quad (10)$$

- CC-AMN underperforms AMN due to numerically instabilities in the optimization landscape:

Table: Mean evaluation score (and standard deviation)

Model	Breakout	Atlantis	Pong	SpaceInv.
DQN	331 (±44)	32.8 (±14.4)	20.9 (±0.2)	442 (±119)
AMN	337 (±74)	31.6 (±9.1)	20.9 (±0.1)	415 (±126)
CC-AMN	320 (±66)	26.2 (±10.4)	8.8 (±11.9)	415 (±132)

Conclusions

- The Continuous-Categorical enables a probabilistic alternative to the Cross-Entropy loss.
- Theoretically appealing but presents unresolved computational challenges in high dimensions.
- Works well in classification with label-smoothing, but poorly in actor-mimic RL.

Acknowledgements

We thank Andres Potapczynski and the anonymous reviewers for helpful conversations, and the Simons Foundation, Sloan Foundation, McKnight Endowment Fund, NSF 1707398, and the Gatsby Charitable Foundation for support.