Implicit Hypothesis Testing and Divergence Preservation in Neural Network Representations

Kadircan Aksoy[1,2] Protim Bhattacharjee[1] Peter Jung[1,2]

[1]German Aerospace Center (DLR), Institute for Space Research [2]Technical University of Berlin

kadircan.aksoy@dlr.de

arXiv:2601.20477v4

Motivation

- Supervised classification is fundamentally a hypothesis testing problem: decide between class-conditional distributions $\{\mathbb{P}(X \mid Y = y)\}_{y\in\mathcal{Y}}$ for any observation $x \in \mathcal{X}$.
- However, it is not known whether cross-entropy training can achieve the performance of statistically optimal hypothesis tests.

Hypothesis Testing

- Two hypotheses; $H_0 : Z \sim P$ vs. $H_1 : Z \sim Q$ with $Q \ll P$.
- A test on n observations, $\phi_n : \mathcal{Z}^n \to [0, 1]$ incurs type-I and type-II errors:

$$\alpha(\phi_n) := \mathbb{E}_P[\phi_n(Z)], \qquad \beta(\phi_n) := \mathbb{E}_Q[1 - \phi_n(Z)]$$

- **Neyman–Pearson:** [Neyman and Pearson, 1933] For fixed α, the uniformly most powerful test is the log-likelihood ratio test:

$$\phi_n^*(z) = \mathbf{1}\left[\log \frac{dQ}{dP}(z) > \tau\right]$$

- **Stein:** [Polyanskiy and Wu, 2016] The optimal type-II error exponent under fixed type-I error is governed by KL divergence:

$$\lim_{\varepsilon\to 0} \lim_{n\to\infty} -\frac{1}{n} \log \beta_n^*(\varepsilon) = D_{\mathrm{KL}}(Q\|P), \quad \text{where } \beta_n^*(\alpha_n) := \inf_{\phi_n:\, \alpha(\phi_n)\le\alpha_n} \beta(\phi_n).$$

Setup and Main Result

- $X_c := \mathbb{P}(X \mid Y = c)$ and $X_{\bar{c}} := \mathbb{P}(X \mid Y \neq c)$ denotes *high-dimensional* class-conditional and complement input distributions.
- The NP-optimal test for class c requires the log-likelihood ratio:

$$\Lambda_c(x) = \log \frac{X_c(x)}{X_{\bar{c}}(x)}$$

- **Our Hypothesis:** A neural network θ learns a discriminative surrogate for Λ_c on the *lower* dimensional logit space.
- $Z_c := \mathbb{P}(Z \mid Y = c)$ with $Z = \theta(X)$, denotes the class-conditional representation distribution and $\eta_c(x) = \mathbb{P}(Y = c \mid X = x)$, the true class posterior. Define:

$$D_{\mathrm{inp}} := \frac{1}{K}\sum_{c=1}^{K} D_{\mathrm{KL}}(X_c\|X_{\bar{c}}) \quad \text{(task ceiling, from DPI)}$$

$$D_\theta := \frac{1}{K}\sum_{c=1}^{K} D_{\mathrm{KL}}(Z_c\|Z_{\bar{c}}) \le D_{\mathrm{inp}} \quad \text{(retained by } \theta\text{)}$$

$$\delta(\theta) := D_{inp} - D_\theta \quad \text{(divergence gap)}$$

$$\epsilon(\theta) := \mathbb{E}_X[D_{\mathrm{KL}}(\eta(X)\|\sigma(\theta(X)))] \ge 0 \quad \text{(posterior mismatch)}$$

Exact KL Preservation (Theorem 3.1): Let $\mathcal{L}(\theta) := -\mathbb{E}_{(X,Y)}[\log \sigma_Y(\theta(X))]$ be the cross-entropy loss and $\theta^* \in \arg\min_\theta \mathcal{L}(\theta)$. Under the following assumptions:
Realizability $\exists\, \tilde{\theta} \in \Theta$ such that $\sigma_c(\tilde{\theta}(x)) = \eta_c(x)$ μ-a.e.
Bounded Posteriors $\exists\, \kappa \in (0, 1/K)$ such that $\eta_c(x) \ge \kappa$ μ-a.e., ensuring $D_{\mathrm{inp}} < \infty$ and $|\Lambda_c(x)| < \infty$
we have:

$$\epsilon(\theta^*) = 0 \quad \text{and} \quad \delta(\theta^*) = 0$$

Moreover, $\theta^*(X)$ is a **sufficient statistic** for $(X_c, X_{\bar{c}})$ for each c:

$$\exists h_c : \mathbb{R}^K \to \mathbb{R} \quad \text{s.t.} \quad \Lambda_c(x) = h_c(\theta^*(x)) \quad \mu\text{-a.e.}$$

Evidence–Error Plane

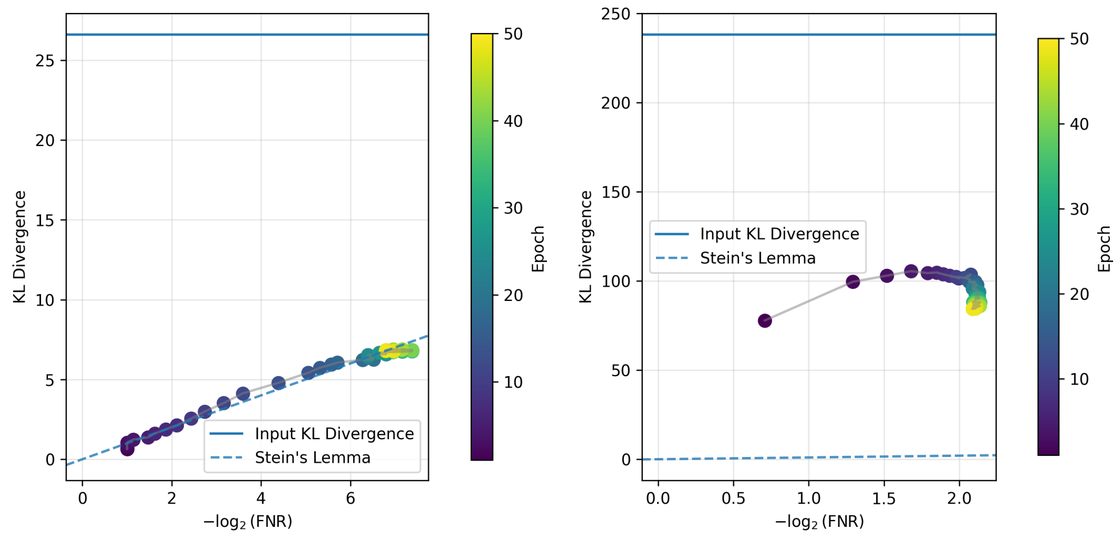
- Define a 2-D coordinate system in which each network is mapped to a point:

$$\Phi(\theta) := (P_\theta,\, D_\theta) \in \mathbb{R}_+^2$$

  where $P_\theta = -\log \beta(\theta)$ is minus log of the *empirical* type-II error.
- The achievable region for all networks is:

$$\mathcal{A} := \{(P_\theta, D_\theta) \in \mathbb{R}_+^2 : 0 \le P_\theta \le D_\theta \le D_{\mathrm{inp}}\}$$

- The diagonal $P_\theta = D_\theta$ is the **Stein limit**. The fully convergent point $(D_{\mathrm{inp}}, D_{\mathrm{inp}})$ corresponds to θ^* in Theorem 3.1.
- Networks with $D_\theta \gg P_\theta$ are **information inefficient**; they extract divergence from the data but can not convert it into performance.
- **Experimentally,** estimate D_θ and D_{inp} with a kNN-based KL divergence estimator (consistent/nearly minimax optimal in MSE for fixed k [Zhao and Lai, 2020] [Pérez-Cruz, 2008]) using test set samples. Similarly calculate empirical error rate $\beta(\theta)$.



Binary Image: 8×8 images corrupted via BSC($p = 0.1$); true D_{inp} analytically known. A fully connected network (64-32-16-8) tracks the Stein limit exactly throughout training.

CIFAR-100: Solved by a ResNet-50. D_{inp} is estimated after a PCA (256 components) due to class imbalance. After epoch 10, D_θ drops while P_θ improves; reminiscent of the compression phase of Information Bottleneck [Shwartz-Ziv and Tishby, 2017], but not predicted by our theory.

Extension to Chernoff Information

- In reality, neural network loss functions minimize type-I and type-II errors simultaneously. The optimal error is characterized by **Chernoff information** [Chernoff, 1952]:

$$C(P, Q) = \sup_{s\in[0,1]} \left[-\log \int_{\mathcal{X}} p(x)^s q(x)^{1-s}\, d\mu(x)\right], \qquad \lim_{n\to\infty} -\frac{1}{n} \log P_e^*(n) = C(P, Q)$$

  where, with priors (π_P, π_Q),

$$P_e^*(n) := \inf_\phi \big(\pi_P \alpha_n(\phi) + \pi_Q \beta_n(\phi)\big)$$

- *A kNN-based estimator for $C(P, Q)$ with L^2 convergence guarantees is in preparation, which will enable evidence–error plane analysis under simultaneous type-I and type-II error minimization.*

References

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231 (694-706):289–337, 1933.

Fernando Pérez-Cruz. Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pages 1666–1670. IEEE, 2008.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. MIT OpenCourseWare, 2016. Course 6.441, Spring 2016.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Puning Zhao and Lifeng Lai. Minimax optimal estimation of kl divergence for continuous distributions. *IEEE Transactions on Information Theory*, 66(12):7787–7811, 2020.