Tailoring Strictly Proper Scoring Rules for Downstream Tasks: An Application to Causal Inference

Roman Plaud* [1,2] Alexandre Perez-Lebel*[3] Antoine Saillenfest [2] Thomas Bonald [1] Marine Le Morvan [4] Gaël Varoquaux [4,5] Matthieu Labeau [1]

[1]Institut Polytechnique de Paris [2]Onepoint [3]Fundamental Technologies [4]INRIA Saclay [5]Probabl

ICML International Conference On Machine Learning

TL;DR

- **The Goal:** A general framework for aligning training losses with downstream estimators geometry.
- **The Method:** Derive a task-specific loss by matching the local curvature of the downstream task error.
- **The Result:** A drop-in loss function that outperforms standard heuristics.

High-level Objective

Loss on Data		Equivalent Objective
Log-Loss	$\Longleftrightarrow$	KL Divergence
$\min_\theta \mathbb{E}\left[\ell_{\text{log-loss}}(Y, p_\theta(X))\right]$		$\min_\theta \mathbb{E}\left[d_{\text{KL}}(p(X), p_\theta(X))\right]$
? ℓ_{task}	$\Longleftrightarrow$	Downstream Task Divergence
$\min_\theta \mathbb{E}\left[?(Y, p_\theta(X))\right]$		$\min_\theta \mathbb{E}\left[d_{\text{task}}(p(X), p_\theta(X))\right]$

Can we reverse-engineer this ? from d_{task}?

Curvature Matching

Task Error Divergence: $d_{\text{task}}(p, p_\theta) \underset{p_\theta \to p}{\approx} \frac{1}{2} \underbrace{w_{\text{task}}(p)}_{\text{Task Curvature}} (p - p_\theta)^2$

Proper Scoring Rule: $d_l(p, p_\theta) \underset{p_\theta \to p}{\approx} \frac{1}{2} \underbrace{w_l(p)}_{\text{Loss Curvature}} (p - p_\theta)^2$

We enforce alignment between the two geometries:

$$w_l(p) = w_{\text{task}}(p)$$

Because every strictly proper scoring rule is uniquely characterized by its weight function, we can analytically recover the exact, optimal training objective $l(Y, p_\theta(X))$.

Application: IPW-Tailored Scoring Rule

Downstream Estimator

IPW for ATE:

$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{Y_i T_i}{p_\theta(X_i)} - \frac{Y_i(1 - T_i)}{1 - p_\theta(X_i)} \right)$$

Bounding the Bias

Isolating propensity:

$$|\text{Bias}|^2 \leq \mathbb{E}\left[d_{\text{task}}(p(X), p_\theta(X))\right]$$

1. Task Divergence:

$$d_{\text{task}}(p, p_\theta) = \left(\frac{p}{p_\theta} - 1 \right)^2 + \left(\frac{1 - p}{1 - p_\theta} - 1 \right)^2$$

2. Task Curvature (second derivative of the divergence):

$$w_{\text{task}}(p) = \frac{2}{p^2} + \frac{2}{(1 - p)^2}$$

3. Loss reconstruction:

$$\ell_{\text{task}}(y, p_\theta) = \frac{2y}{p_\theta} + \frac{2(1-y)}{1-p_\theta} + 2\log(p_\theta(1 - p_\theta))$$

Training Stability: Canonical Mapping

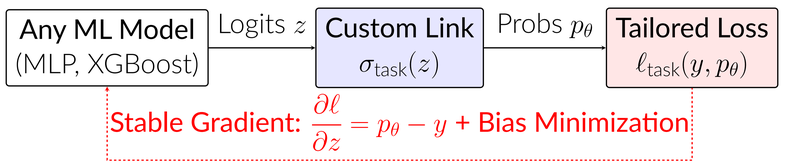
- The Danger: Standard Sigmoid + $\ell_{\text{task}}(p)$ causes gradients to explode near 0 and 1.
- The Fix: We construct a custom link $\sigma_{\text{task}}(z)$ to cancel this curvature by enforcing:

$$(\sigma_{\text{task}}^{-1})' = w_{\text{task}} \iff z p_\theta^2 + (4 - z) p_\theta - 2 = 0$$

yielding the closed-form mapping:

$$\sigma_{\text{task}}(z) = \frac{z - 4 + \sqrt{z^2 + 16}}{2z}$$

A drop-in replacement

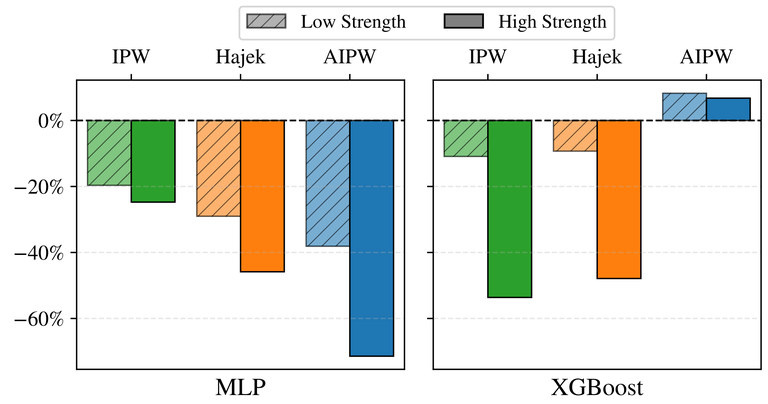


Empirical Results

Method	IPW	Hajek	AIPW
Log-Loss (Baseline)	66.9	8.1	46.5
Custom Loss (Ours)	2.7	2.0	5.1

Across all evaluated datasets, our method is on average the **lowest (best) mean rank**.

We evaluate it as a **drop-in replacement** for log-loss on ACIC 2017 challenge (N=4802, d=58, 32 DGPs):



Take Home Message

- **Ante-Hoc Regularizer:** We mitigate the downstream task error during training.
- **Universal Drop-In:** Replaces log-loss in any standard model (MLPs, XGBoost) with zero architectural changes.
- **Stable:** Paired with its custom canonical link it mathematically guarantees stable gradients.

Paper Code Roman Plaud

 contact: plaud.roman@gmail.com