Calm Composite Losses: Being Improper Yet Proper Composite

Han Bao
Kyoto University

Nontawat Charoenphakdee
Preferred Networks

Abstract

Strict proper losses are fundamental loss functions inducing classifiers capable of estimating class probabilities. While practitioners have devised many loss functions, their properness is often unverified. In this paper, we identify several losses as improper, calling into question the validity of class probability estimates derived from their simplex-projected outputs. Nevertheless, we show that these losses are strictly proper composite with appropriate link functions, allowing predictions to be mapped into true class probabilities. We invent the *calmness* condition, which we prove suffices to identify that a loss has a strictly proper composite representation, and provide the general form of the inverse link. To further understand proper composite losses, we explore proper composite losses through the framework of property elicitation, revealing a connection between inverse link functions and Bregman projections. Numerical simulations are provided to demonstrate the behavior of proper composite losses and the effectiveness of the inverse link function.

1 INTRODUCTION

Loss functions govern numerous aspects of learning problems and their optimal predictors. On the one hand, loss functions define target problems, such as classification by the 0-1 loss, regression by the squared loss (Vapnik, 1998), bipartite ranking by the ranking loss (Freund et al., 2003), classification with abstention by the 0-1-c loss (Herbei and Wegkamp, 2006), and adversarial classification with the robust 0-1 loss (Bao et al., 2020). Once the target loss function is given, one can derive the Bayes optimal predictor for the learning problem. On the other hand, surrogate loss functions are commonly used instead because target loss functions generally entail a discrete nature, hindering direct optimization. Surrogate loss functions exhibit a more optimization-friendly nature, but their optimal predictor could differ from the target optimal predictor. Classification-calibrated losses are loss functions yielding a consistent classifier to the 0-1 loss, and many loss functions are shown to be classification-calibrated, including the logistic and hinge losses (Zhang, 2004; Bartlett et al., 2006; Steinwart, 2007). The gap between target and surrogate loss functions has been more generally studied through the lens of (calibrated) property elicitation (Lambert et al., 2008). In this view, one identifies a loss function as a property (i.e., a multi-valued map from a conditional probability to a report) and studies whether the properties yielded from two loss functions can be linked with each other (Agarwal and Agarwal, 2015). Overall, loss functions dictate what we desire to learn in the form of the optimal predictors and properties—this advocates for the necessity to understand the relationship between a loss function and the corresponding property.

An important class of loss functions is *strictly proper losses*, which induces the true class probability as its optimal predictor (Buja et al., 2005; Gneiting and Raftery, 2007; Reid and Williamson, 2009, 2010; Agarwal, 2014; Williamson et al., 2016; Ovcharov, 2018; Bao, 2023). This is a stronger condition than the classification calibration and enables richer predictors such that they are consistent with AUC (Agarwal, 2014; Menon and Williamson, 2016) and cost-sensitive error rates (Reid and Williamson, 2010; Charoenphakdee et al., 2021a). Indeed, strictly proper losses require a predictor to be as close to the true class probability in L_p distance as possible (Bao, 2023; Bao and Takatsu, 2024). For these reasons, proper losses have been prevailing in risk-sensitive domains. In addition to the standard log loss, we see several examples of proper losses, such as the Brier score and pseudo-spherical score. Proper losses can be transparently understood as the Bregman divergences (Savage, 1971).

On a different track, practitioners have proposed alter-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

native loss functions to the log loss to achieve better robustness. For example, the focal loss has been proposed to stress minority classes (Lin et al., 2017); the generalized cross-entropy loss interpolates the log loss and the mean absolute error (MAE) loss (Zhang and Sabuncu, 2018) for better robustness to symmetric label noises (Ghosh et al., 2017); the inverse focal loss has been proposed to enhance probability calibration with the temperature scaling (Wang et al., 2021). Despite their appealing performances, our understanding of these losses remains elusive when it comes to class probability estimation, except for the recent analysis of the focal loss (Charoenphakdee et al., 2021b). We are interested in whether these loss functions proposed by practitioners are strictly proper. If not, can we rescue improper losses to elicit correct class probabilities?

In this article, we show that all of the above loss functions are improper in themselves but can often be viewed as strictly proper *composite* losses (Williamson et al., 2016). That is, we can recover the true class probability with an appropriate link function that maps an optimal predictor (of the original improper loss) to a true class probability. After reviewing proper losses in Section 2, we show the general form of link functions in Theorem 10 and when a loss function can be turned into a strictly proper composite loss in Section 3. Our main result (Theorem 11) shows that a loss function has a unique, strictly proper composite representation if the loss satisfies the *calmness* condition, which we invent. In Section 4, we show several examples of calm composite losses, which encompass the aforementioned losses. In addition, Section 5 connects composite losses to property elicitation, revealing that the link function is an elicitated property of a given composite loss. Lastly, numerical simulation in Section 6 corroborates that the true conditional probability can be effectively recovered with the link functions.

2 BACKGROUND

We formalize the multi-class learning problem and loss functions in this section. Vectors are denoted in bold face and its i-th element in non-bold face, such as $\mathbf{a}$ and a_i, respectively. The standard inner product of two vectors is denoted by the dual-pair notation $\langle\cdot,\cdot\rangle$. Let $\mathcal{X}$ be an input space and $\mathcal{Y} := [d]$ be a set of labels. The d-probability simplex is denoted by $\triangle^d := \left\{\mathbf{p} \in \mathbb{R}^d_{\geq 0} \mid \langle \mathbf{1}, \mathbf{p}\rangle = 1\right\}$. We identify $\triangle^d$ as the set of all admissible class-conditional probabilities over $\mathcal{Y}$.

Multi-class learning. In the task of class probability estimation (CPE), we aim to estimate the target conditional probability $\mathbb{P}(Y = y|X = \mathbf{x})$ by a predictor $\widehat{\mathbf{p}} : \mathcal{X} \to \triangle^d$. Let $\boldsymbol{\ell} : \triangle^d \to (-\infty, \infty]^d$ be a loss function, assessing the quality of a probabilistic estimate. The discrepancy between the true class probability and a predictor is measured by the (full) risk functional induced by the loss $\boldsymbol{\ell}$:

$$\mathbb{L}(\widehat{\mathbf{p}}) := \mathop{\mathbb{E}}_{(X,Y)}[\ell_Y(\widehat{\mathbf{p}}(X))],$$

where $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is drawn from the underlying probability distribution. The predictor is trained by minimizing $\mathbb{L}$. To discuss loss functions, we focus on the pointwise perspective of the risk functional. First, let us introduce the *conditional risk* associated with $\boldsymbol{\ell}$:

$$L(\mathbf{p}, \widehat{\mathbf{p}}) := \sum_{y\in\mathcal{Y}} p_y \ell_y(\widehat{\mathbf{p}}) = \langle \boldsymbol{\ell}(\widehat{\mathbf{p}}), \mathbf{p}\rangle\,, \tag{1}$$

where $\mathbf{p}, \widehat{\mathbf{p}} \in \triangle^d$ should be understood as the true and estimated probability vectors, respectively. Then, the full risk can be rewritten pointwisely, such as $\mathbb{L}(\mathbf{p}) = \mathbb{E}_X\, L(\mathbb{P}(Y|X), \widehat{\mathbf{p}}(X))$. Therefore, we drop the outer expectation $\mathbb{E}_X$ and focus on the inner (conditional) risk hereafter. The sufficient regularity conditions have been previously discussed; confer Steinwart (2007) and Bao and Takatsu (2024). Correspondingly, we define the *conditional Bayes risk* as the minimal conditional risk at a given true class probability:

$$\underline{L}(\mathbf{p}) := \inf_{\widehat{\mathbf{p}}\in\triangle^d} L(\mathbf{p}, \widehat{\mathbf{p}}) = \inf_{\widehat{\mathbf{p}}\in\triangle^d} \langle \boldsymbol{\ell}(\widehat{\mathbf{p}}), \mathbf{p}\rangle\,.$$

When necessary, we use superscripts to clarify the dependency on the underlying loss, such as $L^{\boldsymbol{\ell}}$ and $\underline{L}^{\boldsymbol{\ell}}$.

Proper loss. The presentation of this part largely owes to Bao and Takatsu (2024); readers unfamiliar with basic convex analysis may refer to it. We first state a regularity condition. A loss $\boldsymbol{\ell} : \triangle^d \to (-\infty, \infty]^d$ is *regular* if $\ell_y(\mathbf{p}) = \infty$ happens only for $y \notin \operatorname{supp}(\mathbf{p})$ (Gneiting and Raftery, 2007).

Definition 1. *A loss* $\boldsymbol{\ell} : \triangle^d \to (-\infty, \infty]^d$ *is* proper *if* $L(\mathbf{p}, \widehat{\mathbf{p}}) \geq L(\mathbf{p}, \mathbf{p}) = \underline{L}(\mathbf{p})$ *for all* $\mathbf{p}, \widehat{\mathbf{p}} \in \triangle^d$*. A loss is* strictly proper *if the inequality is strict for* $\mathbf{p} \neq \widehat{\mathbf{p}}$*.*

For the CPE problem, the properness is a minimal condition to ensure the loss makes sense. As seen, a proper loss is closely related to a convex function.

Proposition 2. *A regular loss* $\boldsymbol{\ell}$ *is (strictly) proper iff there exists a proper (strictly) convex function* Λ *on* $\mathbb{R}^d$ *such that* $\operatorname{dom}(\Lambda) = \triangle^d$ *and, for all* $\widehat{\mathbf{p}} \in \triangle^d$*, there exists a subgradient* $\widehat{\mathbf{v}} \in \partial\Lambda(\widehat{\mathbf{p}})$ *satisfying*

$$L(\mathbf{p}, \widehat{\mathbf{p}}) = -\Lambda(\widehat{\mathbf{p}}) - \langle \widehat{\mathbf{v}}, \mathbf{p} - \widehat{\mathbf{p}}\rangle \quad \text{for all } \mathbf{p} \in \triangle^d.$$

Proposition 3. *For a regular loss* $\boldsymbol{\ell}$*, if we define a proper convex function by*

$$\Lambda(\mathbf{p}) := \begin{cases} -\underline{L}(\mathbf{p}) & \text{if } \mathbf{p} \in \triangle^d \\ \infty & \text{otherwise} \end{cases},$$

then we have $-\boldsymbol{\ell}(\mathbf{p}) \in \partial\Lambda(\mathbf{p})$ *for all* $\mathbf{p} \in \triangle^d$*.*

Propositions 2 and 3 redisplay Proposition 5 and Corollary 6 of Bao and Takatsu (2024), and their ideas date back to Savage (1971, Section 4) and McCarthy (1956, Theorem 1), respectively, with a different way of presentations. With these claims, one can relate a given proper loss to a convex function (by Proposition 2) and generate a proper loss from a convex function (by Proposition 3). We can additionally see that the conditional Bayes risk $\underline{L}$ is concave.

Further, Proposition 2 implies the connection between a proper loss and the Bregman divergence.

Definition 4. *Let $\Lambda : \mathbb{R}^d \to (-\infty, \infty]$ be a proper convex function and $\nabla\Lambda : \triangle^d \to [-\infty, \infty)^d$ be its subdifferential selector. For $\mathbf{p}, \mathbf{p}^0 \in \triangle^d$, the* Bregman divergence *of $\mathbf{p}$ at $\mathbf{p}^0$ is defined by*

$$B_{(\Lambda,\nabla\Lambda)}(\mathbf{p}\|\mathbf{p}^0) \coloneqq \Lambda(\mathbf{p}) - \Lambda(\mathbf{p}^0) - \left\langle \nabla\Lambda(\mathbf{p}^0), \mathbf{p} - \mathbf{p}^0 \right\rangle .$$

For a proper loss $\boldsymbol{\ell}$, define the *(conditional) regret* as the suboptimality of an estimate $\widehat{\mathbf{p}}$ from a true $\mathbf{p}$:

$$R(\mathbf{p}, \widehat{\mathbf{p}}) \coloneqq L(\mathbf{p}, \widehat{\mathbf{p}}) - \underline{L}(\mathbf{p}).$$

We witness $R(\mathbf{p}, \widehat{\mathbf{p}}) = B_{(\Lambda,-\boldsymbol{\ell})}(\mathbf{p}\|\widehat{\mathbf{p}})$ by combining all the above, which indicates that the risk minimization with a proper loss (or equivalently, the conditional regret minimization) is nothing else but the Bregman projection. For example, one can readily see that the log loss is associated with the negative Shannon entropy $\Lambda(\mathbf{p}) = \langle \log \mathbf{p}, \mathbf{p} \rangle$. The relation between a proper loss and the Bregman divergence has been extensively leveraged from the optimization perspective, leading to a new loss family called the *Fenchel–Young losses* (Blondel et al., 2020). The connection between proper losses and Fenchel–Young losses has been discussed in Bao and Sugiyama (2021).

3 PROPER COMPOSITE REPRESENTATION

Even though loss functions like the focal loss are improper per se, they can be reshaped as a proper loss combined with an appropriate link function. The idea has been originated from $\boldsymbol{\Psi}^\gamma$-transform introduced by Charoenphakdee et al. (2021b). We significantly generalize it by using the composite-loss perspective.

3.1 Review of Composite Loss

We briefly recapitulate the composite-loss framework, borrowing the presentation of Williamson et al. (2016). Let $\mathcal{R}$ be a report space, which is an auxiliary prediction space in addition to $\triangle^d$.[1] Instead of predictions working on $\triangle^d$ directly as in Section 2, we consider predictions over $\mathcal{R}$. We use another notation $\boldsymbol{\lambda} : \mathcal{R} \to \mathbb{R}^d_{\geq 0}$ to denote such a loss function. Suppose there exists a continuous invertible function $\boldsymbol{\psi} : \triangle^d \to \mathcal{R}$. Then, one can obtain a final prediction on $\mathbf{p} = \triangle^d$ from an intermediate report $\mathbf{r} \in \mathcal{R}$ by $\mathbf{p} = \boldsymbol{\psi}^{-1}(\mathbf{r})$. With this $\boldsymbol{\psi}$, we can write $\boldsymbol{\lambda} = \boldsymbol{\ell} \circ \boldsymbol{\psi}^{-1}$ with a loss function $\boldsymbol{\ell}$ working on $\triangle^d$. In this regard, we call $\boldsymbol{\psi}$ *link function*, $\boldsymbol{\lambda}$ *composite loss*, and $\boldsymbol{\ell}$ *base loss*. If the base loss $\boldsymbol{\ell}$ is (strictly) proper, $\boldsymbol{\lambda}$ is correspondingly said (strictly) proper composite.

Composite losses are widely used in machine learning. Specifically, for the report space $\mathcal{R} = \mathbb{R}^d$, the cross-entropy loss can be framed as a proper composite loss with the choice of $\boldsymbol{\ell}(\mathbf{p}) = -\log(\mathbf{p})$ and $[\psi^{-1}]_y(\mathbf{r}) = \exp(r_y)/\sum_i \exp(r_i)$. For our purpose, we will take $\mathcal{R} = \triangle^d$. Indeed, previous investigations on composite losses beyond the choice of $\mathcal{R} = \mathbb{R}^d$ have been scarce; our subsequent results are also interesting in their own right to go beyond this choice.

One important question is when $\boldsymbol{\lambda}$ has a proper composite representation $\boldsymbol{\lambda} = \boldsymbol{\ell} \circ \boldsymbol{\psi}^{-1}$ for some $(\boldsymbol{\ell}, \boldsymbol{\psi})$, and its uniqueness. The answers have been provided in Williamson et al. (2016, Section 5.3). To state them, let us introduce basic objects for convex geometry. Note that $\mathcal{R}$ is still free from a specific choice within this subsection.

Definition 5. *Let $h_{\mathbf{p}}^{p_0} \coloneqq \left\{ \mathbf{z} \in \mathbb{R}^d \;\middle|\; \langle \mathbf{z}, \mathbf{p} \rangle = p_0 \right\}$ be a hyperplane. It* supports *a set A at $\mathbf{z} (\in A)$ when $\mathbf{z} \in h_{\mathbf{p}}^{p_0}$ and $\langle \mathbf{z}', \mathbf{p} \rangle \geq p_0$ for all $\mathbf{z}' \in A$.*[2] *For brevity, we say (the normal vector) $\mathbf{p}$* supports *A at $\mathbf{z}$.*

Definition 6. *For a loss $\boldsymbol{\lambda} : \mathcal{R} \to \mathbb{R}^d_{\geq 0}$, the* loss image *$\boldsymbol{\lambda}(\mathcal{R})$ is defined as $\boldsymbol{\lambda}(\mathcal{R}) \coloneqq \{\boldsymbol{\lambda}(\mathbf{r}) \mid \mathbf{r} \in \mathcal{R}\}$.*

Definition 7. *A loss image $\boldsymbol{\lambda}(\mathcal{R})$ is* $\triangle^d$-strictly convex *if for all $\mathbf{p} \in \triangle^d$, there exists a unique $\mathbf{z} \in \boldsymbol{\lambda}(\mathcal{R})$ such that $\mathbf{p}$ supports $\boldsymbol{\lambda}(\mathcal{R})$ at $\mathbf{z}$. Additionally, $\boldsymbol{\lambda}(\mathcal{R})$ is* $\triangle^d$-smooth *if for all $\mathbf{z} \in \boldsymbol{\lambda}(\mathcal{R})$, there exists a unique $\mathbf{p} \in \triangle^d$ such that $\mathbf{p}$ supports $\boldsymbol{\lambda}(\mathcal{R})$ at $\mathbf{z}$.*

Interestingly, a proper composite representation can be analyzed via supporting hyperplanes—this owes to the fact that the Bayes risk is essentially a support function (Williamson, 2014). Now we can state the existence and uniqueness of a proper composite representation, which corresponds to Propositions 18 and 13 from Williamson et al. (2016), respectively.

Proposition 8 (Existence)**.** *Suppose $\boldsymbol{\lambda} : \mathcal{R} \to \mathbb{R}^d_{\geq 0}$ is a continuous loss. Then, $\boldsymbol{\lambda}$ has a proper composite representation iff $\boldsymbol{\lambda}(\mathcal{R})$ is $\triangle^d$-smooth. Moreover, it*

[1] The terminology 'report space' is borrowed from the literature on property elicitation, e.g., Finocchiaro et al. (2019). Therein, an agent is asked to *report* her belief about the probability of an event.

[2] The last condition $\langle \mathbf{z}', \mathbf{p} \rangle \geq p_0$ confirms that the set A lies in the only side of the halfspace defined by $h_{\mathbf{p}}^{p_0}$, and not in the other halfspace $\{\mathbf{z}' \mid \langle \mathbf{z}', \mathbf{p} \rangle < p_0\}$. Nevertheless, the one side suffices for our purpose.

is strictly proper iff $\boldsymbol{\lambda}(\mathcal{R})$ *is additionally* $\triangle^d$*-strictly convex.*

Proposition 9 (Uniqueness)**.** *Suppose that a loss* $\boldsymbol{\lambda} : \mathcal{R} \to \mathbb{R}^d_{\geq 0}$ *has a proper composite representation* $\boldsymbol{\lambda} = \boldsymbol{\ell} \circ \boldsymbol{\psi}^{-1}$*. Then,* $\boldsymbol{\ell}$ *is unique almost everywhere. If* $\boldsymbol{\lambda}$ *is continuous, then* $\boldsymbol{\ell}$ *is unique everywhere. If* $\boldsymbol{\lambda}$ *is invertible, then* $(\boldsymbol{\ell}, \boldsymbol{\psi})$ *is unique.*

We will leverage them to analyze the structure of composite losses. More thorough studies on the loss geometry have been recently developed by Pacheco and Williamson (2023) and Williamson and Cranko (2023).

Remark 1 ($\boldsymbol{\ell}$ vs. $\boldsymbol{\lambda}$)**.** Subsequently, we repeatedly use two symbols, $\boldsymbol{\lambda}$ and $\boldsymbol{\ell}$, to denote loss functions. When $\boldsymbol{\ell}$ is used, it is always defined on $\triangle^d$ and denotes a CPE loss, as introduced in Section 2. When $\boldsymbol{\lambda}$ is used, the loss is defined on a report space $\mathcal{R}$ and refers to a composite loss. The report space $\mathcal{R}$ may differ from $\triangle^d$, such as the cross-entropy loss. Nonetheless, we mostly choose $\mathcal{R} = \triangle^d$ in this paper, where a given loss shall be regarded as either $\boldsymbol{\ell}$ (a base CPE loss) or $\boldsymbol{\lambda}$ (a composite loss) interchangeably. The difference in the notation expresses how we view a given loss.

3.2 Generalize Link Function

Hereafter, we consider a loss $\boldsymbol{\lambda} : \mathcal{R} \to \mathbb{R}^d_{\geq 0}$ with the prediction set $\mathcal{R} = \triangle^d$ and the following assumptions.

Assumption 1. *A loss* $\boldsymbol{\lambda}$ *is* separable*:* $\lambda_y(\mathbf{r})$ *depends solely on* r_y*. In this case, we write* $\lambda_y : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *and* $\lambda_y(r_y)$ *component-wisely, with abuse of notation. Moreover, we assume that* λ_y *is strictly convex and invertible on* $[0, 1]$*, and* $\lambda_y \in C^1((0, 1))$ *for each* $y \in \mathcal{Y}$*.*

Each λ_y is called the partial loss. The separability is known as the *locality* of proper scoring rules (Parry et al., 2012).[3] Focusing on separable losses is convenient in practical design and implementation. By generalizing Charoenphakdee et al. (2021b, Theorem 11), we derive the minimizer of the conditional risk of $\boldsymbol{\lambda}$.

Theorem 10. *For a loss* $\boldsymbol{\lambda}$ *satisfying Assumption 1, for every* $\mathbf{p} \in \triangle^d$*, if the conditional risk* $L^{\boldsymbol{\lambda}}(\mathbf{p}, \cdot)$ *has a minimizer* $\mathbf{r}^* \in \triangle^d$*,* $\mathbf{p}$ *and* $\mathbf{r}^*$ *satisfy the following relationship:*

$$p_y = \frac{[\lambda'_y(r^*_y)]^{-1}}{\sum_{i=1}^d [\lambda'_i(r^*_i)]^{-1}} \quad \text{for all } y \in \mathcal{Y}. \tag{2}$$

Then, we define $\boldsymbol{\Psi}$-transform for a loss $\boldsymbol{\lambda}$:

$$\boldsymbol{\Psi}(\mathbf{r}) := \left[\frac{[\lambda'_1(r_1)]^{-1}}{\sum_{i=1}^d [\lambda'_i(r_i)]^{-1}} \quad \cdots \quad \frac{[\lambda'_d(r_d)]^{-1}}{\sum_{i=1}^d [\lambda'_i(r_i)]^{-1}} \right]^\top. \tag{3}$$

If $\boldsymbol{\lambda}$ is proper by itself (without any proper composite decomposition), then $\mathbf{p} = \boldsymbol{\Psi}(\mathbf{p})$ should hold. Indeed, the log loss $\lambda_y(p_y) = -\log p_y$ yields $p_y = \Psi_y(\mathbf{p})$ (that is, $\boldsymbol{\Psi} = \mathrm{id}_{\triangle^d}$), but the focal loss $\lambda_y(p_y) = -(1-p_y)^\gamma \log p_y$ yields $p_y \neq \Psi_y(\mathbf{p})$. Note that we can view Eq. (2) as an extension of Reid and Williamson (2010, Corollary 12) to the multi-class case. The perspective of improper losses as proper composite losses has already been mentioned by van Erven et al. (2012), for which we give a general characterization. The proof of Theorem 10 is deferred to Appendix A.

Thus, even if a composite loss $\boldsymbol{\lambda}$ is improper, $\boldsymbol{\Psi}$-transform may recover the true probability from the loss minimizer when $\boldsymbol{\Psi}$ is continuously invertible.

3.3 Calm Composite Losses

The $\boldsymbol{\Psi}$-transform may not be continuously invertible in general, preventing $\boldsymbol{\lambda}$ from being strictly proper composite. We aim to understand what conditions make $\boldsymbol{\lambda}$ decomposable as a proper composite loss.

Theorem 11. *For a loss* $\boldsymbol{\lambda}$ *satisfying Assumption 1, there exists a proper composite representation iff* $\lambda_y \in C^2((0, 1))$ *for each* $y \in \mathcal{Y}$*. The representation is always unique for such* $\boldsymbol{\lambda}$*. Moreover, the decomposition is strictly proper if, for each* $y \in \mathcal{Y}$*,*

$$\lambda'_y < 0 \text{ and } \lambda''_y > 0 \text{ on } (0, 1), \text{ and } \lim_{r_y \downarrow 0} \lambda'_y(r_y) = -\infty.$$

We call the last three conditions calmness*, and* $\boldsymbol{\lambda}$ *satisfying them a* calm composite loss*.*[4]

The proof (in Appendix A) shows the bijectivity of $\boldsymbol{\Psi}$.[5] Theorem 11 shows that $\boldsymbol{\lambda}$ is strictly proper composite with a unique decomposition if each partial loss λ_y is twice continuously differentiable and has nondegenerate first- and second-order derivatives with diverging λ'_y at the endpoint 0. This calmness condition is convenient because we can avoid the challenging task of testing the strict properness of the base $\boldsymbol{\ell}$ (with $\boldsymbol{\lambda} = \boldsymbol{\ell} \circ \boldsymbol{\Psi}$)—generally difficult due to the inaccessibility to the analytical form of $\boldsymbol{\ell}$.[6] By checking the calmness, we can know whether $\boldsymbol{\ell}$ associated with $\boldsymbol{\lambda}$ is strictly proper without accessing its analytical form.

[3] In the context of the Bregman divergence, it is more common to say a function is *separable* (Ovcharov, 2018), and thus we opt to use this jargon.

[4] We call it the calmness because the divergent λ'_y calms p_y down to 0 as $r_y \downarrow 0$. Otherwise, $r_y \downarrow 0$ solely does not calm p_y down.

[5] We are aware that Wang and Scott (2024, Lemma G.6) shows the bijectivity of link functions for the other class of multi-class loss functions (called PERM losses) by leveraging an M-matrix, which is a different proof strategy from ours.

[6] When $d = 2$, a loss is strictly proper iff its Bayes risk is strictly concave (Agarwal, 2014, Theorem 4). It requires the analytical form of the Bayes risk, and moreover, we are unaware of a similar characterization for $d > 2$.

Table 1: Loss functions that are calm (✔) and are not (✘). Only the log loss is proper as a base loss among these examples. MAE and GCE indicate the mean absolute error and generalized cross-entropy losses, respectively. "SPC" indicates whether the loss always has (✔) a Strictly Proper Composite loss for any $d \geq 2$ and γ (or N) in the range or does not always have (✘). More discussions of these examples and counterexamples for ✘ can be found in Appendix B.

Loss name	Param. range	$\lambda(r)$	$\lambda'(r)$	Calm	SPC
Log	—	$-\log r$	$-r^{-1}$	✔	✔
Focal	$\gamma \geq 0$	$-(1-r)^{\gamma} \log r$	$(1-r)^{\gamma-1}(\gamma \log r - \frac{1-r}{r})$	✔	✔
Inverse focal	$\gamma \in [0,1]$	$-(1+r)^{\gamma} \log r$	$-(1+r)^{\gamma-1}(\gamma \log r + \frac{1+r}{r})$	✔	✔
GCE	$\gamma \in (0,1)$	$(1-r^{\gamma})/\gamma$	$-r^{\gamma-1}$	✔	✔
MAE	—	$1-r$	-1	✘	✘
Power	$\gamma \geq 1$	$(1-r)^{\gamma}$	$-\gamma(1-r)^{\gamma-1}$	✘	✘
Taylor CE	$N \in \mathbb{N}$	$\sum_{\gamma=1}^{N}(1-r)^{\gamma}/\gamma$	$-\sum_{\gamma=1}^{N}(1-r)^{\gamma-1}$	✘	✘
Inverse focal	$\gamma = 1.5$ and $d = 2$	—	—	✘	✔

Note that the calmness is only *sufficient* for the *strict* properness; nonetheless, the characterization of sufficient conditions for multi-class losses has long been challenging (Saerens et al., 2002, Section IV). Our theorem does not rule out the possibility that a strictly proper composite representation exists for a given loss. Yet, many well-known losses are calm, as in Table 1.

4 EXAMPLES

We look at examples such as the log loss, focal loss (Lin et al., 2017), inverse focal loss (Wang et al., 2021), and generalized cross-entropy (GCE) loss (Zhang and Sabuncu, 2018), and summarize them in Table 1. Non-calm losses are discussed in Appendix B.

4.1 Log Loss

Though the log loss is strictly proper as a base loss, let us recast it as a composite loss $\lambda_y(r_y) = -\log r_y$. Its derivatives are $\lambda'_y(r_y) = -r_y^{-1}$ and $\lambda''_y(r_y) = r_y^{-2}$. Hence, the log loss is a calm composite loss. Its $\mathbf{\Psi}$-transform is $\mathbf{\Psi}(\mathbf{r}) = \mathbf{r}$, and the inverse $\mathbf{\Psi}$ is identity, which means that the base loss $\boldsymbol{\ell}$ is the log loss itself. Thus, we see an example yielding itself as a base loss when interpreted as a composite loss.

4.2 Focal Loss

For the focal loss $\lambda_y(r_y) = -(1-r_y)^{\gamma} \log r_y$, we have

$$\lambda'_y(r_y) = (1-r_y)^{\gamma-1}\left(\gamma \log r_y - \frac{1-r_y}{r_y}\right),$$

$$\lambda''_y(r_y) = (1-r_y)^{\gamma-2}\Bigg(-(\gamma-1)\gamma \log r_y - (2\gamma-1) + \frac{2(\gamma-1)}{r_y} + \frac{1}{r_y^2}\Bigg),$$

from which $\lambda'_y < 0$, $\lambda''_y > 0$, and $\lambda'_y(r_y) \to -\infty$ as $r_y \downarrow 0$ hold. Hence, the focal loss is a calm composite loss. Although $\mathbf{\Psi}$ cannot be analytically inverted, we can guarantee the base loss $\boldsymbol{\ell}$ is strictly proper.

4.3 Inverse Focal Loss

The inverse focal loss $\lambda_y(r_y) = -(1+r_y)^{\gamma} \log r_y$ has been shown empirically to promote model overconfidence, where $\gamma \in (0,3]$ was used in practice. We have

$$\lambda'_y(r_y) = -(1+r_y)^{\gamma-1}\left(\gamma \log r_y + \frac{1+r_y}{r_y}\right),$$

$$\lambda''_y(r_y) = -(1+r_y)^{\gamma-2}\Bigg((\gamma-1)\gamma \log r_y + (2\gamma-1) + \frac{2(\gamma-1)}{r_y} - \frac{1}{r_y^2}\Bigg),$$

where we can verify the calmness. Note that $\gamma > 1$ violates the calmness because the convexity of λ_y is lost—nevertheless, $(\gamma, d) = (1.5, 2)$ yields strictly proper composite at least, demonstrating that the calmness is merely sufficient. We show this counterexample numerically in Fig. 4 (b) of Appendix B.

4.4 Generalized Cross-entropy Loss (GCE)

The GCE loss $\lambda_y(r_y) = (1-r_y^{\gamma})/\gamma$ (with $\gamma \in (0,1)$) has been used in classification from noisy labels. It interpolates the log loss ($\gamma = 0$) and the MAE loss ($\gamma = 1$). Their derivatives are

$$\lambda'_y(r_y) = -r^{\gamma-1} \quad \text{and} \quad \lambda''_y(r_y) = -(\gamma-1)r^{\gamma-2}.$$

We observe that $\lambda'_y < 0$, $\lambda''_y > 0$, and $\lambda'_y(r_y) \to -\infty$ as $r_y \downarrow 0$. Therefore, GCE loss is a calm composite loss.

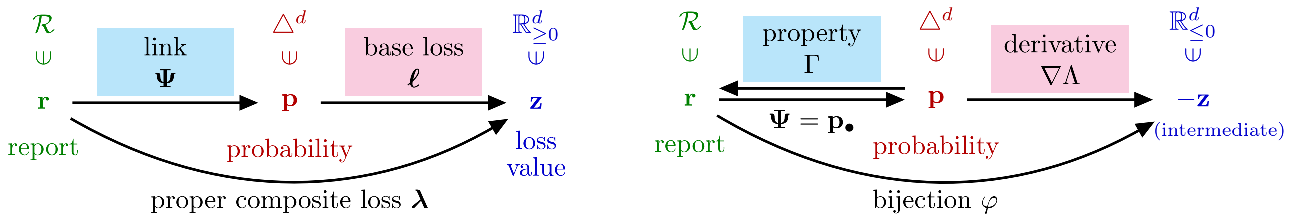


Figure 1: (Left) Composite loss discussed in Section 3, adapted from Williamson et al. (2016, Figure 1). The composite loss $\boldsymbol{\lambda}$ is decomposed into the link $\boldsymbol{\Psi}$ and the base loss $\boldsymbol{\ell}$, which are two knobs. **(Right)** Property elicitation discussed in Section 5. The property Γ is a central object herein, and the loss value $\mathbf{z}$ is merely an intermediate object between a probability and a report. The composite loss $\boldsymbol{\lambda}$ can be understood as (negative of) the bijection φ in property elicitation.

5 ANOTHER VIEW OF $\boldsymbol{\Psi}$ FROM PROPERTY ELICITATION

We have witnessed that an improper loss can have a proper composite representation under the calmness condition. In this section, we further attempt to understand what property a proper composite loss elicits.

5.1 Review of Property Elicitation

We briefly summarize property elicitation, following Frongillo and Kash (2021), with their new connections to proper losses and the Bregman divergence.

Definition 12. *Let $\mathcal{R}$ be a given report space. A* property *is a multi-valued map $\Gamma : \triangle^d \rightrightarrows \mathcal{R}$ which associates a nonempty set of correct report values to each probability $\mathbf{p} \in \triangle^d$. We let $\Gamma^{-1}(\mathbf{r}) := \left\{\mathbf{p} \in \triangle^d \,\middle|\, \mathbf{r} \in \Gamma(\mathbf{p})\right\}$ denote the level set of Γ at $\mathbf{r}$.*[7]

A property is a summary statistic of a probability distribution. For example, mean and variance can be written in the form of a property (Lambert et al., 2008). The central question in property elicitation is to ask what score elicits a property of our interest.

Definition 13. *Let $S : \triangle^d \times \mathcal{R} \to (-\infty, \infty]$ be an affine score, meaning that $S(\cdot, \mathbf{r})$ is affine for every $\mathbf{r} \in \mathcal{R}$. S* elicits *a property $\Gamma : \triangle^d \rightrightarrows \mathcal{R}$ if for all $\mathbf{p} \in \triangle^d$,*[8]

$$\Gamma(\mathbf{p}) = \operatorname*{arg\,min}_{\mathbf{r}\in\mathcal{R}} S(\mathbf{p}, \mathbf{r}).$$

A property Γ is elicitable *if there exists some affine score S eliciting Γ.*

In essence, a property is elicitable if it is the minimizer of S. The conditional risk (1) is an example of an affine score. We define the *Bayes score* of S as follows:

$$\underline{S}(\mathbf{p}) := S(\mathbf{p}, \mathbf{r_p}) \quad \text{where} \quad \mathbf{r_p} \in \Gamma(\mathbf{p}),$$

where any choice of $\mathbf{r_p}$ gives well-defined $\underline{S}$. To characterize affine scores, we need two more regularity conditions.

- A property $\Gamma : \triangle^d \rightrightarrows \mathcal{R}$ is *redundant* if there exists $\mathbf{r}, \mathbf{r}' \in \mathcal{R}$ such that $\Gamma^{-1}(\mathbf{r}') \subsetneq \Gamma^{-1}(\mathbf{r})$. Otherwise, Γ is *non-redundant*.
- An affine score $S : \triangle^d \times \mathcal{R} \to (-\infty, \infty]$ is Γ-*regular* if $S(\mathbf{p}, \mathbf{r}) \in \mathbb{R}$ whenever $\mathbf{r} \in \Gamma(\mathbf{p})$.

If a property is redundant, we can eliminate "redundant" reports (i.e., $\mathbf{r}'$ in the above statement) to make it non-redundant in pre-processing, making subsequent statements simpler.

The following general characterization was established by Frongillo and Kash (2021, Theorem 4.5). In this statement, $\overline{\partial\Lambda} := \bigcup_{\mathbf{p}\in\triangle^d} \partial\Lambda(\mathbf{p})$ denotes the set of all possible subgradients of Λ.

Proposition 14. *Suppose $\Gamma : \triangle^d \rightrightarrows \mathcal{R}$ is non-redundant and $S : \triangle^d \times \mathcal{R} \to (-\infty, \infty]$ is Γ-regular. Then, S elicits Γ iff there exist a convex function $\Lambda : \triangle^d \to \mathbb{R}$, $\mathcal{D} \subseteq \overline{\partial\Lambda}$, a bijection $\varphi : \mathcal{R} \to \mathcal{D}$ with $\Gamma(\mathbf{p}) = \varphi^{-1}(\mathcal{D} \cap \partial\Lambda(\mathbf{p}))$, and a mapping $\mathbf{p_r} \in \Gamma^{-1}(\mathbf{r})$ for each $\mathbf{r} \in \mathcal{R}$ such that for all $\mathbf{p} \in \triangle^d$ and $\mathbf{r} \in \mathcal{R}$,*

$$S(\mathbf{p}, \mathbf{r}) = -\Lambda(\mathbf{p_r}) - \langle \varphi(\mathbf{r}), \mathbf{p} - \mathbf{p_r} \rangle.$$

Proposition 14 contains several technicalities to deal with the "inverse" of Γ, yet its message is much simpler. Compare its statement with Proposition 2, which characterizes the admissible form of the conditional risk of a proper loss. The conditional regret of a proper loss $\boldsymbol{\ell}$ can be written as the Bregman divergence:

$$\begin{aligned} L^{\boldsymbol{\ell}}(\mathbf{p}, \widehat{\mathbf{p}}) - \underline{L}^{\boldsymbol{\ell}}(\mathbf{p}) &= \Lambda(\mathbf{p}) - \Lambda(\widehat{\mathbf{p}}) - \langle -\boldsymbol{\ell}(\widehat{\mathbf{p}}), \mathbf{p} - \widehat{\mathbf{p}} \rangle \\ &= B_{(\Lambda, -\boldsymbol{\ell})}(\mathbf{p} \| \widehat{\mathbf{p}}), \end{aligned}$$

where Λ is chosen in Proposition 3. Similarly, Proposition 14 states that the conditional regret satisfies

$$\begin{aligned} S(\mathbf{p}, \mathbf{r}) - \underline{S}(\mathbf{p}) &= \Lambda(\mathbf{p}) - \Lambda(\mathbf{p_r}) - \langle \varphi(\widetilde{\Gamma}(\mathbf{p_r})), \mathbf{p} - \mathbf{p_r} \rangle \\ &= B_{(\Lambda, \varphi\circ\widetilde{\Gamma})}(\mathbf{p} \| \mathbf{p_r}), \end{aligned}$$

[7] Note that Γ^{-1} is the inverse image but not the inverse map of Γ.

[8] We consider the score minimization instead of the maximization as in Frongillo and Kash (2021) to make the following presentation consistent with Proposition 2.

where $\widetilde{\Gamma}$ is a single-valued version of Γ chosen by $\widetilde{\Gamma}(\mathbf{p}_r) = \mathbf{r}$ for each $\mathbf{r} \in \mathcal{R}$, through which we can interpret a property Γ as the Bregman projection of $\mathbf{p}$ to the report space $\mathcal{R}$. This gives us an information-geometric understanding of property elicitation.

Remark 2. In the above, $\varphi \circ \widetilde{\Gamma}$ behaves as a subdifferential selector of Λ. Such another composed property $\varphi \circ \Gamma$ is said *subgradient-elicitable* because this property is directly represented as a subgradient of some convex function. The subgradient elicitability is stronger than the standard elicitability. To get back to proper losses, a (negative base) proper loss $-\boldsymbol{\ell}$ is subgradient-elicitable (with the convex function $\Lambda = -\underline{L}$). This is a translation of Proposition 3 into the jargon of property elicitation. More discussions on subgradient elicitation can be found in Frongillo and Kash (2021, Section 5.1).

5.2 Elicitation via Composite Loss

We saw that a (base) proper loss $-\boldsymbol{\ell}$ itself is elicited by the associated conditional risk $L^{\boldsymbol{\ell}}$ as an affine score. Then, what does the conditional risk $L^{\boldsymbol{\lambda}}$ associated with the composite loss $\boldsymbol{\lambda}$ elicit? Here, we suppose that $\boldsymbol{\lambda}$ is calm and has a strictly proper composite representation $\boldsymbol{\lambda} = \boldsymbol{\ell} \circ \boldsymbol{\Psi}$ (with the $\boldsymbol{\Psi}$-transform (3) being identified as the inverse link $\boldsymbol{\psi}^{-1}$).

In light of Proposition 14, let us identify the conditional risk $L^{\boldsymbol{\lambda}} : \triangle^d \times \mathcal{R} \to (-\infty, \infty]$ as an affine score, and the inverse link $\boldsymbol{\Psi} : \mathcal{R} \to \triangle^d$ as $\mathbf{p}_\bullet : \mathcal{R} \to \triangle^d$, where the report space $\mathcal{R} = \triangle^d$. First, we derive the convex function Λ corresponding to the Bayes score.

Lemma 15. *If a loss $\boldsymbol{\lambda}$ has a proper composite representation $\boldsymbol{\lambda} = \boldsymbol{\ell} \circ \boldsymbol{\psi}^{-1}$, $\underline{L}^{\boldsymbol{\lambda}} = \underline{L}^{\boldsymbol{\ell}}$ holds.*

Hence, we can identify $\Lambda = -\underline{L}^{\boldsymbol{\lambda}} = -\underline{L}^{\boldsymbol{\ell}}$. We omit the proof because it immediately follows from the definitions of the properness and conditional risk.

Next, we derive the bijection φ of Proposition 14. When $\boldsymbol{\lambda}$ is calm, $\Lambda(= -\underline{L}^{\boldsymbol{\lambda}})$ is differentiable and $\partial\Lambda(\mathbf{p}) = \{-\boldsymbol{\ell}(\mathbf{p})\}$ by Proposition 3. Then, by $\Gamma(\mathbf{p}) = \varphi^{-1}(\mathcal{D} \cap \partial\Lambda(\mathbf{p}))$, Γ is single-valued and invertible. With the choice of $\mathbf{p}_\bullet = \boldsymbol{\Psi}$, the inverse of the property $\Gamma^{-1} = \boldsymbol{\Psi}$, where Γ^{-1} denotes the inverse map of the single-valued map Γ. Thus,

$$\varphi(\mathbf{r}) = -\boldsymbol{\ell} \circ \Gamma^{-1}(\mathbf{r}) = -\boldsymbol{\ell} \circ \boldsymbol{\Psi}(\mathbf{r}) \quad \text{for all } \mathbf{r} \in \mathcal{R}.$$

All in all, the formulas $\varphi = -\boldsymbol{\ell} \circ \boldsymbol{\Psi}$ and $\varphi \circ \Gamma = -\boldsymbol{\ell}$ are established, and the target property is nothing else but $\boldsymbol{\Psi}^{-1}$. This implies that the conditional regret of the composite loss $\boldsymbol{\lambda}$ is

$$L^{\boldsymbol{\lambda}}(\mathbf{p}, \mathbf{r}) - \underline{L}^{\boldsymbol{\lambda}}(\mathbf{p}) = B_{(\Lambda, -\boldsymbol{\ell})}(\mathbf{p} \| \boldsymbol{\Psi}(\mathbf{r}))$$

for all $\mathbf{p}, \mathbf{r} \in \triangle^d$. Now we see that $L^{\boldsymbol{\lambda}}$ elicits the link $\boldsymbol{\Psi}^{-1}$ via the Bregman projection with $B_{(\Lambda, -\boldsymbol{\ell})}$.

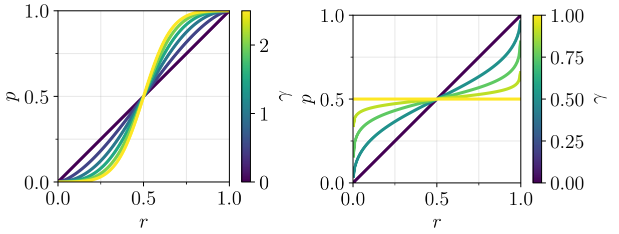


Figure 2: Illustration of $\boldsymbol{\Psi}$-transform for the binary case. **(Left)** Focal loss with $\gamma \in \{0, 0.5, \ldots, 2.5\}$. The log loss ($\gamma = 0$) yields $\Psi = \mathrm{id}$. **(Right)** GCE loss with $\gamma \in \{0, 0.5, 0.75, 0.9, 1\}$. The log loss ($\gamma = 0$) yields $\Psi = \mathrm{id}$, while the MAE loss ($\gamma = 1$) yields $\Psi \equiv 0.5$.

Here, the target property and the geometry of the Bregman projection are determined by the inverse link $\boldsymbol{\Psi}$ and base loss $\boldsymbol{\ell}$, respectively. This is another interpretation of the 'two tuning knobs' (see Section 3.3) of a proper composite loss. The structure discussed here is summarized in Fig. 1.

5.3 Illustration of Property

To understand the target property Γ more, we illustrate its inverse $\boldsymbol{\Psi}$ in the binary case for clarity. We reparametrize $\mathbf{r} = [r_1 \ r_2]^\top \in \triangle^2$ by a single scalar $r \in [0, 1]$ by $\mathbf{r} = [r \ 1-r]^\top$ and consider

$$p = \Psi(r) := \frac{[\lambda_1'(r)]^{-1}}{[\lambda_1'(r)]^{-1} + [\lambda_2'(1-r)]^{-1}} \in [0, 1],$$

corresponding to the $\boldsymbol{\Psi}$-transform of class 1.

Figure 2 shows $p = \Psi(r)$ for the focal and GCE losses. For the focal loss, $\gamma = 0$ (log loss) entails the diagonal Ψ, meaning that $\Gamma = \mathrm{id}$ and the log loss is proper (as a base CPE loss). For $\gamma > 0$, Ψ is distorted so that

$$\left|p - \tfrac{1}{2}\right| > \left|r - \tfrac{1}{2}\right|,$$

which means that the reported value r is always less "confident" (i.e., close to the binary decision boundary 1/2). This nature of the focal loss may be useful when a user wants to prevent over-confident predictions. Note that the under-confident nature of the focal loss is specific to $d = 2$; it becomes under- or over-confident adaptively to the input report in general $d > 2$ (Charoenphakdee et al., 2021b, Section 4.2).

For the GCE loss, $\gamma = 1$ entails $\Psi \equiv 0.5$, which aligns with the fact that the MAE loss ($\gamma = 1$) is not strictly proper. As long as $\gamma \in [0, 1)$, the GCE loss is calm, and its Ψ encourages over-confident reports. However, as $\gamma \uparrow 1$, Ψ becomes less steep and closer to a constant link. Indeed, $\gamma = 0.9$ yields a near-flat Ψ, though it is still surjective, indicating that the GCE loss with γ closer to 1 performs less effectively in CPE numerically.

In summary, the target property associated with a composite loss $\boldsymbol{\lambda}$ "distorts" a given probability (Fig. 2).

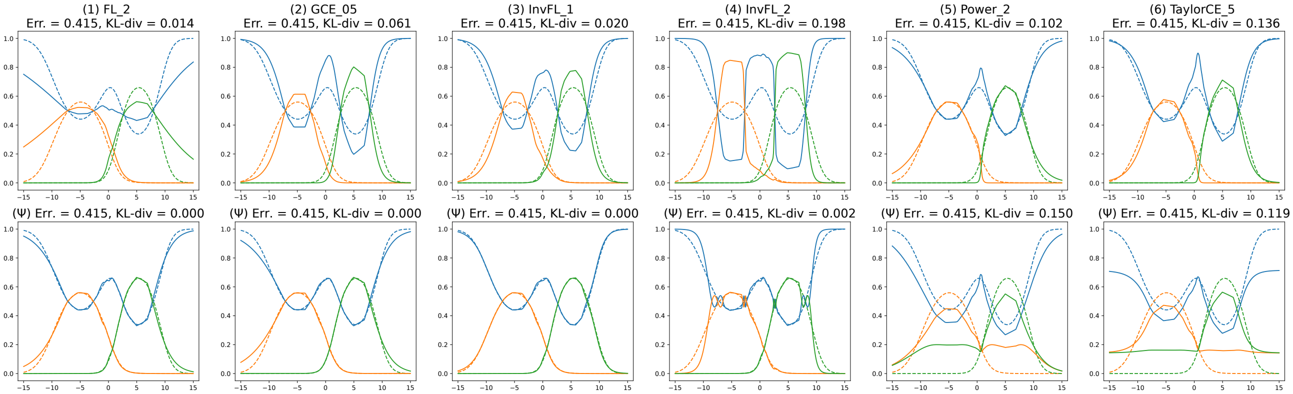


Figure 3: Numerical simulation results: the dashed lines represent the ground truth class probabilities, while the solid lines show the estimated probabilities. The first row displays the results using the softmax outputs of the MLPs, and the second row presents the results after applying the **Ψ**-transform. "Err." denotes the classification error, and "KL-div." refers to the KL divergence between the estimated and true class probabilities. Here, only the focal loss, GCE loss, inverse focal with $\gamma = 1$ are calm composite.

This practically matters when one does not trust the underlying probability; in this case, she can modulate the target property to obtain a distorted report.

Remark 3. When the target probability $\mathbf{p}$ is contaminated with noise, Sypherd et al. (2022) discussed how to recover the true probability as a report $\mathbf{r}$ by leveraging the distortion nature of the target property. They introduced a notion called *twist-properness* of a loss family $\{\boldsymbol{\lambda}^\gamma\}_\gamma$ parametrized by γ, which informally means that there exists γ for every possible noisy channel such that $\boldsymbol{\lambda}^\gamma$ recovers the correct probability. While the twist-properness cares about a loss *family*, the calmness cares about a loss *instance*.

6 SIMULATION

We provide numerical simulation to showcase how **Ψ**-transform recovers the class probabilities from the softmax report. We followed the experimental setting of the numerical simulation conducted in Charoenphakdee et al. (2021b, Section 5.2) by considering a three-class classification problem, where the feature is one-dimensional generated from three different univariate Gaussian distributions: $(\mu_1, \sigma_1) = (0, 4)$, $(\mu_2, \sigma_2) = (-3, 2.5)$, and $(\mu_3, \sigma_3) = (4, 2)$. The input distribution $p(x)$ is a mixture of three Gaussian distributions weighted by $(w_1, w_2, w_3) = (2, 1, 1)$. The number of training data is 100,000. We trained a 3-layer MLP with 128 hidden nodes in each layer. LeakyReLU with the slope of -0.2 was used. The full-batch gradient descent was used for training with 20,000 epochs, the learning rate 0.005, and momentum 0.9.

We compared six following losses: (1) focal loss (FL) ($\gamma = 2$), (2) GCE ($\gamma = 0.5$), (3) inverse focal loss (InvFL) ($\gamma = 1$), (4) inverse focal loss ($\gamma = 2$), (5) Taylor cross-entropy loss (TaylorCE) ($N = 5$), and (6) power loss ($\gamma = 2$). Note that only losses (1)–(3) are ensured to be strictly proper composite by Theorem 11.

Figure 3 illustrates the performance of each loss function, showing (1)–(6) from left to right. We used KL-divergence (KL-div) as a metric to evaluate the quality of class probability estimation. The first row presents the softmax outputs, where we observe that no losses accurately estimate probabilities, suggesting they are improper. It can also be observed that the classification error (Err.) is similar for all losses.

The second row shows the probability estimates obtained after applying the **Ψ**-transform derived from Theorem 10. In this case, only losses (1)–(3) successfully recover the true class probabilities, while losses (4)–(6) exhibit significantly worse performance. The worse performance can be seen by the recovered class probabilities (denoted in the solid lines) and the KL divergence values. It is worth noting that the inverse focal loss with $\gamma = 2$ is no longer a convex loss, and hence Theorem 11 is not applicable. Yet, it does not perform as effectively as the inverse focal loss with $\gamma = 1$. Overall, the simulation result agrees with our theoretical results. We provide full results with more different choices of γ and N for all losses in Appendix C.

7 CONCLUSION

We provided a perspective to view improper CPE losses as strictly proper composite losses, for which we identified that the calmness condition suffices. Checking whether a separable loss is calm is fairly straightforward and is amenable to practical usage. Thus, we can recover the true class probability from the minimizer of a calm loss via the **Ψ**-transform.

Acknowledgment

HB is supported by JST PRESTO (Grant No. JPMJPR24K6). We would like to appreciate Nan Lu for giving a pointer to van Erven et al. (2012) and Jayakorn Vongkulbhisal for providing suggestions on the experiment setup and issues related to the power loss. Also, we would like to thank an anonymous reviewer to point out a relevant paper, Saerens et al. (2002).

References

Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of the 28th Conference on Learning Theory*, pages 4–22, 2015.

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

Han Bao. Proper losses, moduli of convexity, and surrogate regret bounds. In *Proceedings of the 36th Conference on Learning Theory*, pages 525–547, 2023.

Han Bao and Masashi Sugiyama. Fenchel-Young losses with skewed entropies for class-posterior probability estimation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 1648–1656, 2021.

Han Bao and Asuka Takatsu. Proper losses regret at least 1/2-order. *arXiv preprint arXiv:2407.10417*, 2024.

Han Bao, Clayton Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Proceedings of the 33rd Conference on Learning Theory*, pages 408–451, 2020.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Anders Björner, Jiří Matoušek, and Günter M Ziegler. Using Brouwer's fixed point theorem. *A Journey Through Discrete Mathematics: A Tribute to Jiří Matoušek*, pages 221–271, 2017.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Technical Report*, 2005.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *Proceedings of the 36th International Conference on Machine Learning*, pages 961–970, 2019.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1507–1517, 2021a.

Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, and Masashi Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5202–5211, 2021b.

Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, pages 2206–2212, 2021.

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. *Advances in Neural Information Processing Systems*, 32:10781–10791, 2019.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

Rafael M Frongillo and Ian A Kash. General truthfulness characterizations via convex analysis. *Games and Economic Behavior*, 130:636–662, 2021.

Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102 (477):359–378, 2007.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

Nicolas S Lambert, David M Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42 (9):654–655, 1956.

Aditya Krishna Menon and Robert C Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

Evgeni Y Ovcharov. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1):53–79, 2018.

Armando J Cabrera Pacheco and Robert Williamson. The geometry of mixability. *Transactions on Machine Learning Research*, 2023.

Mattew Parry, A Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

Mark D Reid and Robert C Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th International Conference on Machine Learning*, pages 897–904, 2009.

Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Any reasonable cost function can be used for a posteriori probability approximation. *IEEE Transactions on Neural Networks*, 13(5):1204–1210, 2002.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Tyler Sypherd, Richard Nock, and Lalitha Sankar. Being properly improper. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20891–20932, 2022.

Tim van Erven, Mark D Reid, and Robert C Williamson. Mixability is Bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, 13:1639–1663, 2012.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820, 2021.

Yutong Wang and Clayton Scott. Unified binary and multiclass margin-based classification. *Journal of Machine Learning Research*, 25(143):1–51, 2024.

Robert C Williamson. The geometry of losses. In *Proceedings of the 27th Conference on Learning Theory*, pages 1078–1108, 2014.

Robert C Williamson and Zac Cranko. The geometry and calculus of losses. *Journal of Machine Learning Research*, 24(342):1–72, 2023.

Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31:8778–8788, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sections 2 and 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] We do not propose a new algorithm.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No] Our simulation remains synthetic and not challenging to replicate.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 3.
   (b) Complete proofs of all theoretical results. [Yes] See Appendix A.
   (c) Clear explanations of any assumptions. [Yes] See Assumption 1 and Theorem 11.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The dataset detail is described in Section 6.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 6.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [No] We mainly focus on the visual presentation of the estimated results in Fig. 3, where we do not add error bars to avoid visual clutters.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No] Our simulation is synthetic, so we do not need a large-scale cluster to replicate.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOFS

Proof of Theorem 10. The Lagrangian associated with the minimization problem of $L^{\boldsymbol{\lambda}}(\mathbf{p},\cdot)$ is written as follows:

$$\mathcal{L}(\mathbf{p},\beta) \coloneqq \sum_{y\in\mathcal{Y}} p_y \lambda_y(r_y) + \beta\left(\sum_{y\in\mathcal{Y}} r_y - 1\right),$$

where $\beta \geq 0$ is the Lagrangian multiplier. From the KKT conditions, we have the following identity at the minimum $\mathbf{r} = \mathbf{r}^*$:

$$\left.\frac{\partial \mathcal{L}}{\partial r_y}\right|_{\mathbf{r}=\mathbf{r}^*} = p_y \lambda_y'(r_y^*) + \beta = 0 \quad \text{for all } y \in \mathcal{Y}.$$

By plugging this into the normalization condition $\sum_y p_y = 1$, we have

$$\beta = -\frac{1}{\sum_{y\in\mathcal{Y}} \frac{1}{\lambda_y'(r_y^*)}}.$$

From this expression of β, Eq. (2) can be verified. □

Proof of Theorem 11. We first show the **existence of proper composite representation**. To see this, we need to show $\boldsymbol{\lambda}(\triangle^d)$ is $\triangle^d$-smooth by Proposition 8. In other words, for any $\mathbf{z} \in \boldsymbol{\lambda}(\triangle^d)$, we verify whether there exists a unique $\mathbf{p} \in \triangle^d$ such that a hyperplane $h_{\mathbf{p}}^{p_0}$ (see Definition 5) supports $\boldsymbol{\lambda}(\triangle^d)$ at $\mathbf{z}$. When $\mathbf{p}$ supports $\boldsymbol{\lambda}(\triangle^d)$ at $\mathbf{z} \in \boldsymbol{\lambda}(\triangle^d)$, $\langle \mathbf{z}', \mathbf{p}\rangle \geq \langle \mathbf{z}, \mathbf{p}\rangle$ holds for all $\mathbf{z}' \in \boldsymbol{\lambda}(\triangle^d)$, meaning that the following holds:

$$\underbrace{\inf_{\mathbf{z}'\in\boldsymbol{\lambda}(\triangle^d)} \langle \mathbf{z}', \mathbf{p}\rangle}_{=:\widehat{\sigma}_{\boldsymbol{\lambda}(\triangle^d)}(\mathbf{p})} = \langle \mathbf{z}, \mathbf{p}\rangle\,, \tag{4}$$

where $\widehat{\sigma}_A$ is the support function of a set A, slightly different from the standard support function defined by the supremum—interested readers may refer to Williamson (2014, Section 2) for the distinctions in detail. Since each λ_y is invertible (Assumption 1), there exists an inverse $\mathbf{r}$ with $r_y = \lambda_y^{-1}(z_y)$ for each $y \in \mathcal{Y}$. Similarly, an inverse $\mathbf{r}'$ exists with $r_y' = \lambda_y^{-1}(z_y')$ for each $y \in \mathcal{Y}$. Then, Eq. (4) is equivalently rewritten as follows:

$$\underbrace{\min_{\mathbf{r}'\in\triangle^d} \langle \boldsymbol{\lambda}(\mathbf{r}'), \mathbf{p}\rangle}_{=\underline{L}^{\boldsymbol{\lambda}}(\mathbf{p})} = \langle \boldsymbol{\lambda}(\mathbf{r}), \mathbf{p}\rangle\,.$$

The minimizer is attained at the condition $\mathbf{p} = \boldsymbol{\Psi}(\mathbf{r})$ by Theorem 10. Hence, the infimum of Eq. (4) with respect to $\mathbf{z}$ is attained by

$$\mathbf{p} = \boldsymbol{\Psi}(\boldsymbol{\lambda}^{-1}(\mathbf{z})) \eqqcolon \mathbf{f}(\mathbf{z}), \tag{5}$$

where $\mathbf{f} \coloneqq \boldsymbol{\Psi} \circ \boldsymbol{\lambda}^{-1}$. Actually, $\mathbf{f}$ is well-defined because of the invertibility assumption of $\boldsymbol{\lambda}$. The uniqueness of $\mathbf{p}$ given an arbitrary $\mathbf{z}$ is translated into the differentiability of $\mathbf{f}$ at every $\mathbf{z} \in \boldsymbol{\lambda}(\mathrm{int}(\triangle^d))$. To confirm this, we calculate the Jacobian of $\mathbf{f}$. Noting that we have only $(d-1)$ "degrees of freedom" in $\mathbf{f}$ because of $\sum_{i=1}^d r_i = \sum_{i=1}^d \lambda_i^{-1}(z_i) = 1$, we confine $\mathbf{f}$ by

$$\overline{\mathbf{f}}(z_1,\ldots,z_{d-1}) \coloneqq \begin{bmatrix} \overline{f}_1(z_1,\ldots,z_{d-1}) \\ \vdots \\ \overline{f}_{d-1}(z_1,\ldots,z_{d-1}) \end{bmatrix} \quad \text{and}$$

$$\overline{f}_y(z_1,\ldots,z_{d-1}) \coloneqq f_y(z_1,\ldots,z_{d-1},z_d) = \frac{[\lambda_y' \circ \lambda_y^{-1}(z_y)]^{-1}}{\sum_{i=1}^{d-1}[\lambda_i' \circ \lambda_i^{-1}(z_i)]^{-1} + \left[\lambda_d'\left(1 - \sum_{j=1}^{d-1}\lambda_j^{-1}(z_j)\right)\right]^{-1}},$$

where z_d is determined by $\sum_{j=1}^d r_j = 1$ such that

$$z_d = \lambda_d\left(1 - \sum_{j=1}^{d-1} \lambda_j^{-1}(z_j)\right),$$

and calculate the Jacobian of $\overline{\mathbf{f}}$. If $k \neq y$ (for $k, y \in \mathcal{Y} \setminus \{d\}$),

$$
\begin{aligned}
\frac{\partial \overline{f}_y}{\partial z_k} &= -\frac{[\lambda_y' \circ \lambda_y^{-1}(z_y)]^{-1}}{\left\{\sum_{i=1}^d [\lambda_i' \circ \lambda_i^{-1}(z_i)]^{-1}\right\}^2} \cdot \frac{\partial}{\partial z_k}\left\{[\lambda_k' \circ \lambda_k^{-1}(z_k)]^{-1} + \left[\lambda_d'\left(1 - \sum_{j=1}^{d-1}\lambda_j^{-1}(z_j)\right)\right]^{-1}\right\} \\
&= -\frac{[\lambda_y' \circ \lambda_y^{-1}(z_y)]^{-1}}{\left\{\sum_{i=1}^d [\lambda_i' \circ \lambda_i^{-1}(z_i)]^{-1}\right\}^2} \cdot \left\{-\frac{\lambda_k'' \circ \lambda_k^{-1}(z_k)}{\lambda_k' \circ \lambda_k^{-1}(z_k)} + \frac{\lambda_d'' \circ \lambda_d^{-1}(z_d)}{[\lambda_d' \circ \lambda_d^{-1}(z_d)]^2} \cdot \frac{1}{\lambda_k' \circ \lambda_k^{-1}(z_k)}\right\} \\
&= \overline{f}_y(\overline{\mathbf{z}}) \cdot \overline{f}_k(\overline{\mathbf{z}}) \cdot \left\{\lambda_k'' \circ \lambda_k^{-1}(z_k) - \frac{\lambda_d'' \circ \lambda_d^{-1}(z_d)}{[\lambda_d' \circ \lambda_d^{-1}(z_d)]^2}\right\},
\end{aligned}
$$

where we write $\overline{\mathbf{z}} := [z_1 \ \dots \ z_{d-1}]^\top \in \mathbb{R}_{\geq 0}^{d-1}$ for brevity. If $k = y$ (for $y \in \mathcal{Y} \setminus \{d\}$), we have similarly

$$
\frac{\partial \overline{f}_y}{\partial z_y} = [\overline{f}_y(\overline{\mathbf{z}})]^2 \cdot \left\{\lambda_y'' \circ \lambda_y^{-1}(z_y) - \frac{\lambda_d'' \circ \lambda_d^{-1}(z_d)}{[\lambda_d' \circ \lambda_d^{-1}(z_d)]^2}\right\} - \overline{f}_y(\overline{\mathbf{z}}) \cdot [\lambda_y'' \circ \lambda_y^{-1}(z_y)].
$$

Combining them, the Jacobian of $\overline{\mathbf{f}}$ (for $\overline{\mathbf{z}}$ such that $\mathbf{z} \in \boldsymbol{\lambda}(\mathrm{int}(\triangle^d))$) is calculated as follows:

$$
\mathrm{D}\overline{\mathbf{f}}(\overline{\mathbf{z}}) = \underbrace{\begin{bmatrix} \vdots \\ \overline{f}_y(\overline{\mathbf{z}}) \\ \vdots \end{bmatrix}}_{=:\mathbf{u}} \underbrace{\begin{bmatrix} \cdots & \overline{f}_y(\overline{\mathbf{z}}) \cdot \left\{\lambda_y'' \circ \lambda_y^{-1}(z_y) - \frac{\lambda_d'' \circ \lambda_d^{-1}(z_d)}{[\lambda_d' \circ \lambda_d^{-1}(z_d)]^2}\right\} & \cdots \end{bmatrix}^\top}_{=:\mathbf{v}^\top} - \mathrm{diag}\underbrace{\left(\cdots \ \overline{f}_y(\overline{\mathbf{z}}) \cdot [\lambda_y'' \circ \lambda_y^{-1}(z_y)] \ \cdots\right)}_{=:\mathbf{w}}.
$$

We see that the Jacobian is well-defined under the conditions $\lambda_y' < 0$ and $\lambda_y'' > 0$, whence $\overline{\mathbf{f}}$ is differentiable. Hence, $\boldsymbol{\lambda}(\triangle^d)$ is $\triangle^d$-smooth.

The **uniqueness of the proper composite representation** immediately follows from Assumption 1 and Proposition 9.

To see the **existence of strictly proper composite representation**, we need to show $\boldsymbol{\lambda}(\triangle^d)$ is $\triangle^d$-strictly convex by Proposition 8. In other words, for any $\mathbf{p} \in \triangle^d$, we verify whether there exists a unique $\mathbf{z} \in \boldsymbol{\lambda}(\triangle^d)$ such that a hyperplane $h_{\mathbf{p}}^{\mathrm{po}}$ supports $\boldsymbol{\lambda}(\triangle^d)$ at $\mathbf{z}$. To see the uniqueness of $\mathbf{z}$ given $\mathbf{p}$, we show that $\overline{\mathbf{f}} : (z_1, \dots, z_{d-1}) \mapsto (p_1, \dots, p_{d-1})$ is bijective and hence invertible, which implies the existence of the globally continuous inverse $\mathbf{f}^{-1} : \mathbf{p} \mapsto \mathbf{z}$.

Let us look at the injectivity of $\overline{\mathbf{f}}$. This can be verified if the Jacobian of $\overline{\mathbf{f}}$ is always (strictly) definite. Then, let us calculate $\det(\mathrm{D}\overline{\mathbf{f}}(\overline{\mathbf{z}}))$ by the matrix determinant lemma as follows:

$$
\begin{aligned}
\det(\mathrm{D}\overline{\mathbf{f}}(\overline{\mathbf{z}})) &= \det(\mathbf{u}\mathbf{v}^\top - \mathrm{diag}(\mathbf{w})) \\
&= (1 - \mathbf{v}^\top \mathrm{diag}(\mathbf{w})^{-1}\mathbf{u}) \det(-\mathrm{diag}(\mathbf{w})) \\
&= -\left\{\underbrace{1 - \sum_{y=1}^{d-1} \overline{f}_y(\overline{\mathbf{z}})}_{=\overline{f}_d(\overline{\mathbf{z}})>0} + \underbrace{\frac{\lambda_d'' \circ \lambda_d^{-1}(z_d)}{[\lambda_d' \circ \lambda_d^{-1}(z_d)]^2}}_{>0} \cdot \sum_{y=1}^{d-1} \overline{f}_y(\mathbf{z}) \cdot \underbrace{[\lambda_y'' \circ \lambda_y^{-1}(z_y)]^{-1}}_{>0}\right\} \cdot \underbrace{\det(\mathrm{diag}(\mathbf{w}))}_{>0} \\
&< 0,
\end{aligned}
$$

where the inequalities hold under the conditions $\mathbf{z} \in \boldsymbol{\lambda}(\mathrm{int}(\triangle^d))$, $\lambda_y' < 0$, and $\lambda_y'' > 0$. Thus, $\mathrm{D}\overline{\mathbf{f}}(\overline{\mathbf{z}})$ is always negative definite, and hence, $\overline{\mathbf{f}}$ is injective.

Next, let us check the surjectivity of $\overline{\mathbf{f}}$. Therein, we invoke the next lemma, which is essentially a consequence of Brouwer's fixed point theorem.

Lemma 16 (Björner et al. (2017, Lemma 3.3))**.** *Let P be a convex polytope and let $h : P \to P$ be a continuous map satisfying $h(P) \subseteq P$ for each face F of P. Then, h is surjective.*

We now apply Lemma 16 to $\mathbf{\Psi}$ and conclude that $\mathbf{f}$ defined via Eq. (5) (and $\overline{\mathbf{f}}$) is surjective. $\mathbf{\Psi}$ is a continuous map defined over a convex polytope $\mathcal{R} = \triangle^d$, and ranges over the same convex polytope $\triangle^d$. The continuity is confirmed by the expression (3) and the assumption $\lambda_y \in C^2$ for each $y \in \mathcal{Y}$. Next, fix an arbitrary face of $\triangle^d$ by

$$P \coloneqq \left\{\mathbf{p} \in \triangle^d \;\middle|\; p_1 = 0\right\}$$

without loss of generality. To see $\mathbf{\Psi}(P) \subseteq P$, choose an arbitrary $\mathbf{r} \in P$ and check $\mathbf{\Psi}(\mathbf{r}) \in P$. By the expression of $\mathbf{\Psi}$ in Eq. (3),

$$\Psi_1(\mathbf{r}) = \frac{[\lambda_1'(r_1)]^{-1}}{\sum_{i=1}^d [\lambda_i'(r_i)]^{-1}} = \frac{0}{0 + \sum_{i=2}^d [\lambda_i'(r_i)]^{-1}} = 0,$$

where we use the assumption $\lim_{r_1 \downarrow 0} \lambda_1'(r_1) = -\infty$ ($\implies \lim_{r_1 \downarrow 0} [\lambda_1'(r_1)]^{-1} = 0$). This suggests $\mathbf{\Psi}(\mathbf{r}) \in P$. Indeed, this argument does not depend on the specific choice $y = 1$, and we can establish the same argument for each face of $\triangle^d$. Thanks to Lemma 16, we can assert that $\mathbf{\Psi}$ is surjective. Since we assume each λ_y is invertible, $\mathbf{f} = \mathbf{\Psi} \circ \boldsymbol{\lambda}^{-1}$ (and $\overline{\mathbf{f}}$) is again surjective.

Therefore, $\overline{\mathbf{f}}$ is bijective and has a continuous inverse, which associates a unique $\mathbf{z}$ for all $\mathbf{p} \in \triangle^d$. This asserts that $\boldsymbol{\lambda}(\triangle^d)$ is $\triangle^d$-strictly convex. □

B MORE EXAMPLES OF LOSS FUNCTIONS

Here, we show the MAE loss, power loss, and Taylor cross-entropy loss as examples for which we cannot apply Theorem 11 to verify the existence of strictly proper composite representations.

Example 1 (Mean absolute error (MAE) loss)**.** The mean absolute error (MAE) loss was proposed as a robust loss against class-conditional label noises. It takes the following form, up to a constant factor (Ghosh et al., 2017):

$$\lambda_y(r_y) = 1 - r_y.$$

This loss does not satisfy Assumption 1 as it is a linear function, and therefore, it is not strictly convex.

Example 2 (Power loss)**.** The power loss takes the following form:

$$\lambda_y(r_y) = (1 - r_y)^\gamma,$$

for $\gamma \geq 1$. We generalize the power loss mentioned by Buja et al. (2005) to $d \geq 3$. The power loss satisfies Assumption 1, and the derivatives are

$$\lambda_y'(r_y) = -\gamma(1 - r_y)^{\gamma-1} \quad \text{and} \quad \lambda_y''(r_y) = \gamma(\gamma - 1)(1 - r_y)^{\gamma-2},$$

which satisfies $\lambda_y' < 0$ and $\lambda_y'' > 0$ except for $\gamma = 1$. However, it can be observed that $\lambda_y'(r_y) \to -\gamma > -\infty$ as $r_y \downarrow 0$, which conflicts with the calmness condition. This implies that the power loss cannot be applied to our theoretical result.

Example 3 (Taylor cross-entropy (CE) loss)**.** This loss function was originally proposed for classification with noisy labels (Feng et al., 2021). It takes the following form:

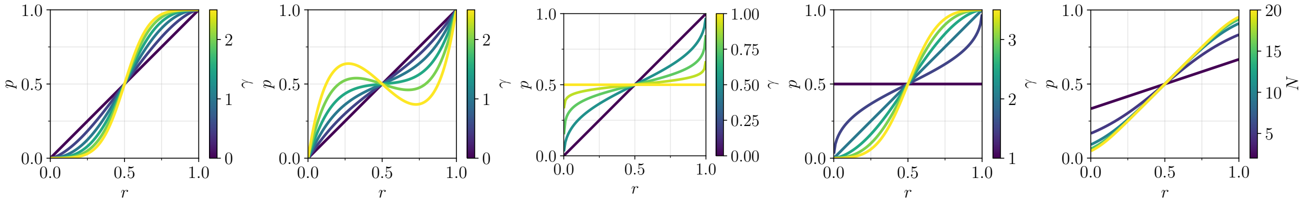
$$\lambda_y(r_y) = \sum_{\gamma=1}^{N} \frac{(1 - r_y)^\gamma}{\gamma},$$

where N is a positive integer.

It can be viewed as the mean absolute error loss when $N = 1$ and cross-entropy loss when $N \to \infty$. Therefore, it can be viewed as yet another interpolation between the mean absolute error and cross-entropy loss, apart from the GCE loss. Their derivatives are

$$\lambda_y'(r_y) = -\sum_{\gamma=1}^{N} (1 - r_y)^{\gamma-1} \quad \text{and} \quad \lambda_y''(r_y) = \sum_{\gamma=1}^{N} (\gamma - 1)(1 - r_y)^{\gamma-2}.$$

We can see that $N = 1$ violates the second derivative condition because $\lambda_y''(r_y) = 0$. For $N > 1$, both conditions $\lambda_y'(r_y) < 0$ and $\lambda_y''(r_y) < 0$ are satisfied. However, it can be observed that $\lambda_y'(r_y) \to -N$ as $r_y \downarrow 0$, which conflicts with the calmness condition. As a result, Taylor CE loss cannot guarantee whether it is a proper composite loss based on our theoretical result.



(a) Focal loss **(b)** Inverse focal loss **(c)** GCE loss **(d)** Power loss **(e)** Taylor CE loss

Figure 4: Illustration of $\mathbf{\Psi}$-transform for the binary case, expanding Fig. 2. Confer Section 5.3 for the illustration details. **(a) Focal loss** with $\gamma \in \{0, 0.5, 1, 1.5, 2, 2.5\}$. The $\mathbf{\Psi}$-transforms are always invertible. **(b) Inverse focal loss** with $\gamma \in \{0, 0.5, 1., 1.5, 2, 2.5\}$. The $\mathbf{\Psi}$-transforms remain invertible up until $\gamma = 1.5$ but are no longer invertible from $\gamma = 2$. **(c) GCE loss** with $\gamma \in \{0, 0.5, 0.75, 0.9, 1\}$. The $\mathbf{\Psi}$-transforms are invertible except for $\gamma = 1$ (MAE loss). **(d) Power loss** with $\gamma \in \{1, 1.5, 2, 2.5, 3, 3.5\}$. The $\mathbf{\Psi}$-transforms are invertible except for $\gamma = 1$ (MAE loss). **(e) Taylor CE loss** with $N \in \{2, 5, 10, 15, 20\}$. The $\mathbf{\Psi}$-transforms are never surjective for finite N, but the range of $\mathbf{\Psi}$ approaches $[0, 1]$ as N increases (where the Taylor CE loss approaches the log loss).

Numerical verification of strict properness ($d = 2$). Although the above three examples cannot be immediately determined whether they have strictly proper composite representation or not, let us provide numerical evidence. In Fig. 4, we expand Fig. 2 in Section 5.3 to illustrate the binary $\mathbf{\Psi}$-transform for more loss functions: focal loss, inverse focal loss, GCE loss, power loss, and Taylor CE loss. As we saw, the focal loss, the inverse focal loss with $\gamma \in [0, 1]$, and the GCE loss with $\gamma \in [0, 1)$ can be guaranteed to be strictly proper composite by Theorem 11, and their $\mathbf{\Psi}$-transforms are invertible.

For the inverse focal loss with $\gamma > 1$, where Theorem 11 is not applicable because the partial loss λ_y is not convex anymore, $\mathbf{\Psi}$ remains invertible for $\gamma = 1.5$, but eventually becomes non-invertible beyond $\gamma \geq 2$. This indicates that the inverse focal loss with $\gamma \geq 2$ does not have any strictly proper composite representation (at least in the case of $d = 2$).

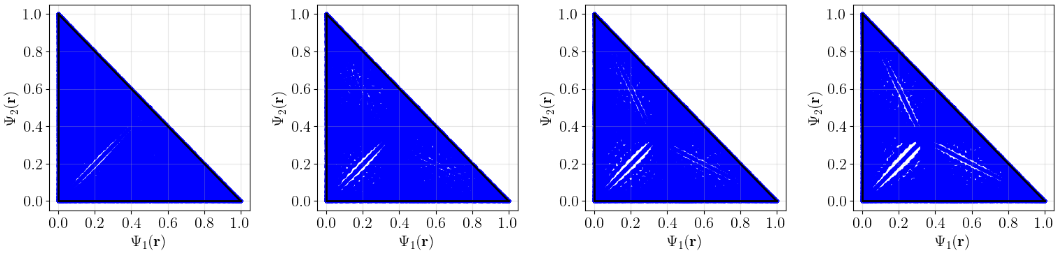
For the power loss, the $\mathbf{\Psi}$-transforms are always invertible when $\gamma > 1$, which suggests that the power loss is also strictly proper composite, at least in the case of $d = 2$. When $\gamma = 1$, the power loss reduces to the MAE loss, and its $\mathbf{\Psi}$ becomes constant, which indicates that the MAE loss does not have any strictly proper composite representation.

For the Taylor CE loss, the $\mathbf{\Psi}$-transforms are always monotonic (injective) but not surjective. Indeed, their ranges are strict subsets of $[0, 1]$. This behavior comes from $\lambda_y'(r_y) \to -N > -\infty$ (as $r_y \downarrow 0$). Thus, the Taylor CE loss does not have any strictly proper composite representation for finite N.

Numerical verification of strict properness ($d = 3$). Beyond the binary case, even the numerical verification of the strict properness becomes challenging. Here, we simply check whether the $\mathbf{\Psi}$-transform of each loss function is surjective or not for the case $d = 3$. If $\mathbf{\Psi}$ is not surjective, then the loss cannot be $\triangle^3$-strictly convex (Definition 7), and does not have a strictly proper composite representation (Proposition 8). In this numerical illustration, report points $\mathbf{r} \in \triangle^3$ are uniformly generated from $h([0, 1]) \times h([0, 1]) \times h([0, 1])$, where the interval $[0, 1]$ each dimension is divided into 40 grids and the distortion $h(x) = x^2$ is applied after the uniform sampling to locate more points near 0. Then, generated report points $\mathbf{r} \in \triangle^3$ are mapped into $\mathbf{p} = \mathbf{\Psi}(\mathbf{r})$, and we numerically plot $[\Psi_1(\mathbf{r}), \Psi_2(\mathbf{r})]$ in the two-dimensional plane.

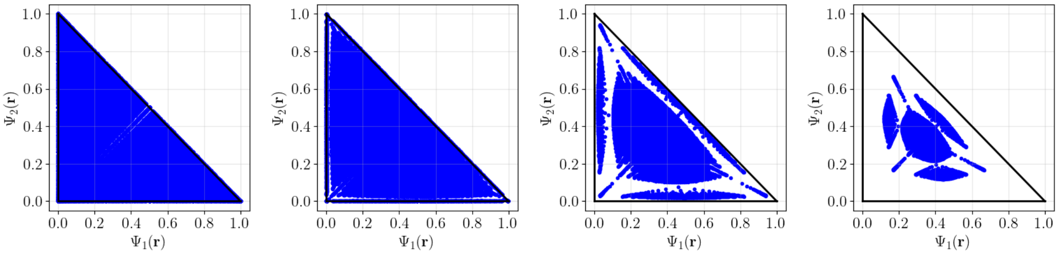
The simulation result is shown in Fig. 5 (the inverse focal loss), Fig. 6 (the GCE loss), Fig. 7 (the power loss), and Fig. 8 (the Taylor CE loss). Figures 5 to 8 should be interpreted by how much the blue points—the uniformly sampled points distorted by $\mathbf{\Phi}$—cover $\triangle^2$. If the entire $\triangle^2$ is filled with the blue points, it indicates $\mathrm{Im}(\mathbf{\Psi}) = \triangle^3$, namely, the surjectivity of $\mathbf{\Psi}$. Among these examples, the inverse focal loss with $\gamma \leq 1$ and the CGE loss with $\gamma < 1$ are shown to have strictly proper composite representations by Theorem 11. Indeed, the inverse focal loss with $\gamma \leq 1$ yields $\mathbf{\Psi}$ whose range covers the entire $\triangle^3$. The range of the GCE link does not appropriately cover the entire $\triangle^3$, especially with $\gamma = 0.90$. This is actually due to numerical errors; as seen in Fig. 4 (c), $\gamma = 0.90$ yields the almost flat $\mathbf{\Psi} \approx 0.5$, which means that most of the input reports are concentrated in a small output range by applying $\mathbf{\Psi}$. Hence, while the GCE loss with γ close to 1 has a strictly proper composite representation in theory, its practical behavior is not necessarily nice.

For the rest of the examples, the power losses and Taylor CE losses, shown in Fig. 7 and Fig. 8, respectively,



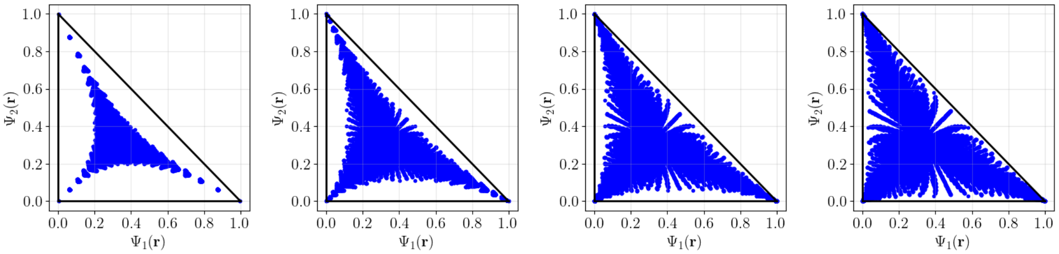
(a) $\gamma = 1.0$ **(b)** $\gamma = 1.5$ **(c)** $\gamma = 2.0$ **(d)** $\gamma = 2.5$

Figure 5: The range of $\mathbf{\Psi}$-transform of the inverse focal loss (guaranteed to be strictly proper composite for $\gamma \leq 1$ by Theorem 11). We observe $\mathrm{Im}(\mathbf{\Psi}) = \triangle^3$ in all cases.



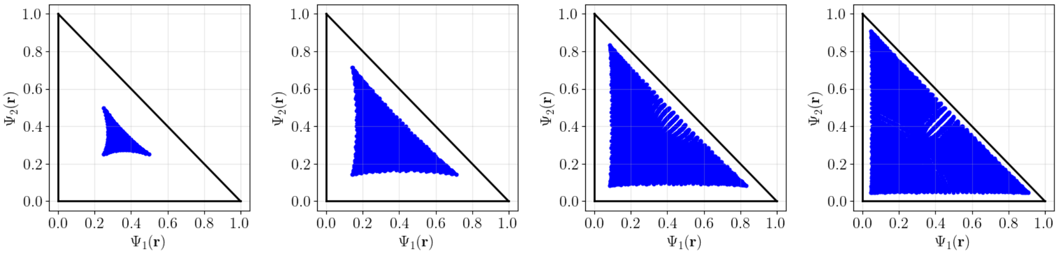
(a) $\gamma = 0.25$ **(b)** $\gamma = 0.50$ **(c)** $\gamma = 0.75$ **(d)** $\gamma = 0.90$

Figure 6: The range of $\mathbf{\Psi}$-transform of the GCE loss (guaranteed to be strictly proper composite by Theorem 11). In theory, $\mathrm{Im}(\mathbf{\Psi}) = \triangle^3$ holds in all cases, but the results are not as good as expected for large γ such as 0.90 due to numerical errors.



(a) $\gamma = 2$ **(b)** $\gamma = 3$ **(c)** $\gamma = 4$ **(d)** $\gamma = 5$

Figure 7: The range of $\mathbf{\Psi}$-transform of the power loss (for which Theorem 11 is not applicable). We observe that $\mathrm{Im}(\mathbf{\Psi}) \subsetneq \triangle^3$ in all cases.



(a) $N = 2$ **(b)** $N = 5$ **(c)** $N = 10$ **(d)** $N = 20$

Figure 8: The range of $\mathbf{\Psi}$-transform of the Taylor CE loss (for which Theorem 11 is not applicable). We observe that $\mathrm{Im}(\mathbf{\Psi}) \subsetneq \triangle^3$ in all cases.

notably fail to cover the entire $\triangle^3$ by their corresponding $\mathbf{\Psi}$. This corroborates the fact that the power and Taylor CE losses do not have strictly proper composite representations. Interestingly, we observe that $\mathrm{Im}(\mathbf{\Psi})$ of the Taylor CE loss shown in Fig. 8 gradually expands as N increases, which aligns well with the fact that the Taylor CE loss converges to the log loss as $N \to \infty$.

The inverse focal losses with $\gamma > 1$ shown in Fig. 5 seem to satisfy $\mathrm{Im}(\mathbf{\Psi}) = \triangle^3$ numerically, but this observation does not immediately assure that they have strictly proper composite representations because we have verified the surjectivity of $\mathbf{\Psi}$ only. In the case of $d = 2$, we can see that $\mathbf{\Psi}$ is not injective with $\gamma \geq 2$, as seen in Fig. 4 (b). Checking the injectivity for $d \geq 3$ is not easy.

Counterexamples for some losses that do not have strictly proper composite representation. Though we do not fully characterize *necessary* and sufficient conditions for a loss to have a strictly proper composite representation in Theorem 11, we can disprove the strict properness for specific examples of loss functions manually by giving counterexamples without relying on the calmness condition.

For the MAE loss, we cannot derive the $\mathbf{\Psi}$-transform by Theorem 10 because the MAE loss is not strictly convex, violating Assumption 1. Instead, we focus on the case $d = 2$ and see

$$\lambda_1(r_1) = 1 - r_1 \quad \text{and} \quad \lambda_2(r_2) = 1 - r_2 = 1 - (1 - r_1) = r_1,$$

which satisfies the *symmetric condition* $\lambda_1(r_1) + \lambda_2(r_2) = \text{const}$. In this case, we can easily show that the true class probability p_1 and the optimal report r_1 of the MAE loss is in the relationship of $r_1 = \mathrm{sign}(p_1 - 1/2)$, by extending Charoenphakdee et al. (2019, Theorem 8). This indicates that the $\mathbf{\Psi}$-transform is not continuously invertible and that the MAE loss for $d = 2$ does not have a strictly proper composite representation.

For the power loss, let us focus on $d = 3$ and $\gamma = 2$:

$$\lambda_y(r_y) = (1 - r_y)^2, \quad \lambda'_y(r_y) = 2(r_y - 1), \quad \text{and} \quad \Psi_y(\mathbf{r}) = \frac{\frac{1}{1-r_y}}{\sum_{i=1}^{3} \frac{1}{1-r_i}}.$$

Now, we confirm that there does not exist $\mathbf{r} \in \triangle^3$ such that $\mathbf{\Psi}(\mathbf{r}) = [1/2 \; 1/2 \; 0]^\top$. The condition $\Psi_3(\mathbf{r}) = 0$ can be written as

$$\frac{\frac{1}{r_3-1}}{\frac{1}{r_1-1} + \frac{1}{r_2-1} + \frac{1}{r_3-1}} = 0,$$

which holds when $r_1 = 1$ or $r_2 = 1$. Note that $1/(r_3 - 1) < 0$ for $\mathbf{r} \in \triangle^3$. However, $r_1 = 1$ and $r_2 = 1$ give $\mathbf{\Psi}(\mathbf{r}) = [1 \; 0 \; 0]^\top$ and $\mathbf{\Psi}(\mathbf{r}) = [0 \; 1 \; 0]^\top$, respectively. Thus, $\mathbf{\Psi}(\mathbf{r}) = [1/2 \; 1/2 \; 0]^\top$ cannot be attained for any $\mathbf{r} \in \triangle^3$, which implies that $\mathbf{\Psi}$ is not surjective in this case.

For the Taylor CE loss, let us focus on $d = 2$ and $N = 2$:

$$\lambda_y(r_y) = 1 - r_y + \frac{(1 - r_y)^2}{2} \quad \text{and} \quad \lambda'_y(r_y) = r_y - 2.$$

By applying Theorem 10,

$$\Psi_1([r_1 \; 1 - r_1]^\top) = \frac{1/(r_1 - 2)}{1/(r_1 - 2) + 1/((1 - r_1) - 2)} = \frac{r_1 + 1}{3}.$$

This indicates that $\mathrm{Im}(\Psi_1) = [1/3, 2/3] \subsetneq [0, 1]$, and $\mathrm{Im}(\mathbf{\Psi}) = \triangle^2$ will never hold. Hence, the Taylor CE loss does not have a strictly proper composite representation in this case.

C ADDITIONAL RESULTS NUMERICAL SIMULATION

In this section, we provide numerical simulation to showcase the effectiveness of Theorem 10 to recover class probabilities from the softmax report. We also provide the results of post-hoc calibration using temperature scaling.

Focal loss. We compared the performance of the focal loss using $\gamma \in \{0.5, 1, 2, 3, 4\}$. Figure 9 shows the performance of the class probability estimation of the focal loss using (1) softmax prediction, (2) $\boldsymbol{\Psi}$-transform, (3) post-hoc temperature scaling of the softmax prediction. It can be observed that the performance of probability estimation of softmax prediction and temperature scaling worsens as γ increases. Nevertheless, $\boldsymbol{\Psi}$ can effectively estimate class probabilities to achieve similar performance for all γ. Our result agrees with the result in Charoenphakdee et al. (2021b).

Inverse focal loss. We compared the performance of the inverse focal loss using $\gamma \in \{0.5, 1, 1.5, 2, 3\}$. Figure 10 shows the performance of the class probability estimation of the inverse focal loss using (1) softmax prediction, (2) $\boldsymbol{\Psi}$-transform, (3) post-hoc temperature scaling of the softmax prediction. It can be observed that the performance of probability estimation of all methods worsens as γ increases after $\gamma > 1.5$. Note that the inverse focal loss is no longer convex when $\gamma > 1$; nevertheless, we can see good performance when $\gamma = 1.5$. However, as $\gamma \geq 2$, the performance degraded. This aligns well with the fact that $\gamma = 1.5$ still gives an invertible $\boldsymbol{\Psi}$, but $\gamma \geq 2$ does not, as seen in Fig. 4 (b).

GCE loss. We compared the performance of the GCE loss using $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Figure 11 shows the performance of the class probability estimation of the GCE loss using (1) softmax prediction, (2) $\boldsymbol{\Psi}$-transform, (3) post-hoc temperature scaling of the softmax prediction. It can be observed that the performance of probability estimation of all methods worsened as γ increases. For the $\boldsymbol{\Psi}$-transform, the performance becomes worse in $\gamma \in \{0.7, 0.9\}$. Even though GCE is strictly proper composite according to Theorem 11, it may not perform well in practice for γ too close to 1. This aligns well with the numerical plots of $\boldsymbol{\Psi}$ in Fig. 4 (c) and Fig. 6. It is worth noting that if $\gamma = 1$, the GCE loss becomes MAE loss, where the $\boldsymbol{\Psi}$-transform becomes a constant function. Therefore, it is also intuitive to think that when $\boldsymbol{\Psi}$-transform is very steep (as when γ is close to 1), the practical performance will be negatively affected.

Power loss. We compared the performance of the power loss using $\gamma \in \{2, 3, 4, 5, 6\}$ respectively. Figure 12 shows the performance of the class probability estimation of the power loss using (1) softmax prediction, (2) $\boldsymbol{\Psi}$-transform, (3) post-hoc temperature scaling of the softmax prediction. Although it does not satisfy Theorem 11, it can be observed that the performance becomes better as γ increases. This may be consistent with the fact that $\mathrm{Im}(\boldsymbol{\Psi})$ approaches the entire $\triangle^d$ with larger γ numerically, as seen in Fig. 7. Nevertheless, its performance (such as KL-div) is still not as good as using the $\boldsymbol{\Psi}$-transform for the focal and inverse focal $(0 < \gamma < 1)$ losses.

Taylor CE loss. We compared the performance of the Taylor CE loss using $N \in \{2, 3, 5, 10, 30\}$. Figure 12 shows the performance of the class probability estimation of the Taylor CE loss using (1) softmax prediction, (2) $\boldsymbol{\Psi}$-transform, (3) post-hoc temperature scaling of the softmax prediction. It can be observed that the performance becomes better as N increases, although it does not satisfy Theorem 11. When N is large, using $\boldsymbol{\Psi}$-transform achieves the worst performance. Note that when $N \to \infty$, the Taylor CE loss is equivalent to the log loss and is equivalent to the MAE loss when $N = 1$.

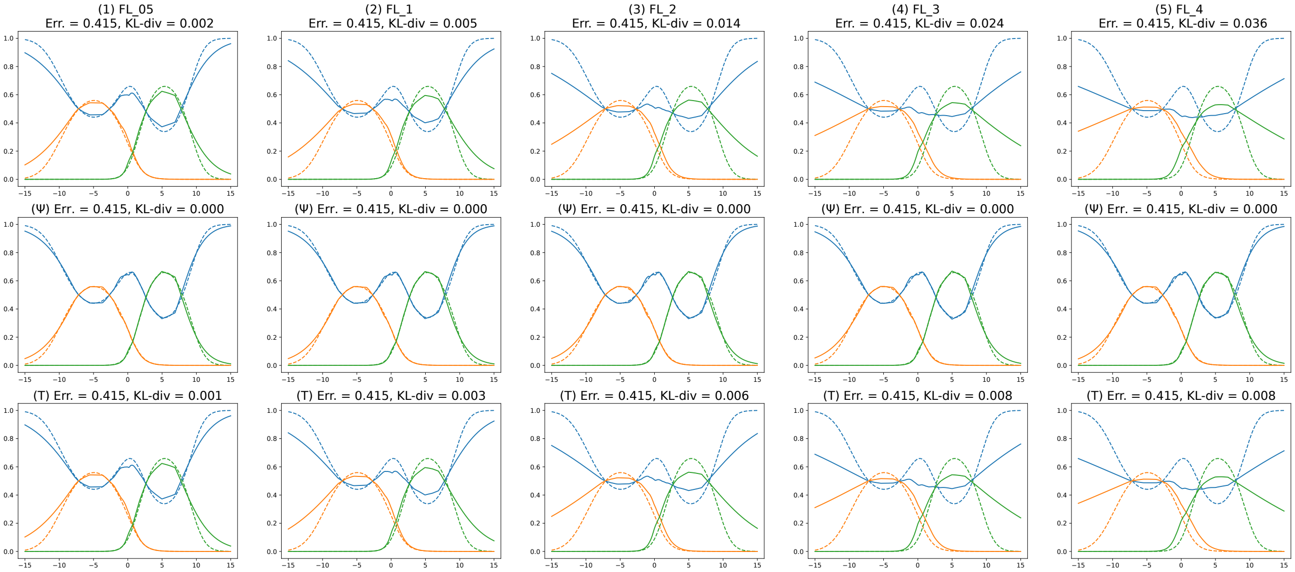


Figure 9: Numerical simulation results of the focal loss, where $\gamma \in \{0.5, 1, 2, 3, 4\}$ from left to right, respectively. The first row indicates the performance of softmax estimation. The second row indicates the performance of $\mathbf{\Psi}$-transform. The third row indicates the performance of post-hoc temperature scaling.

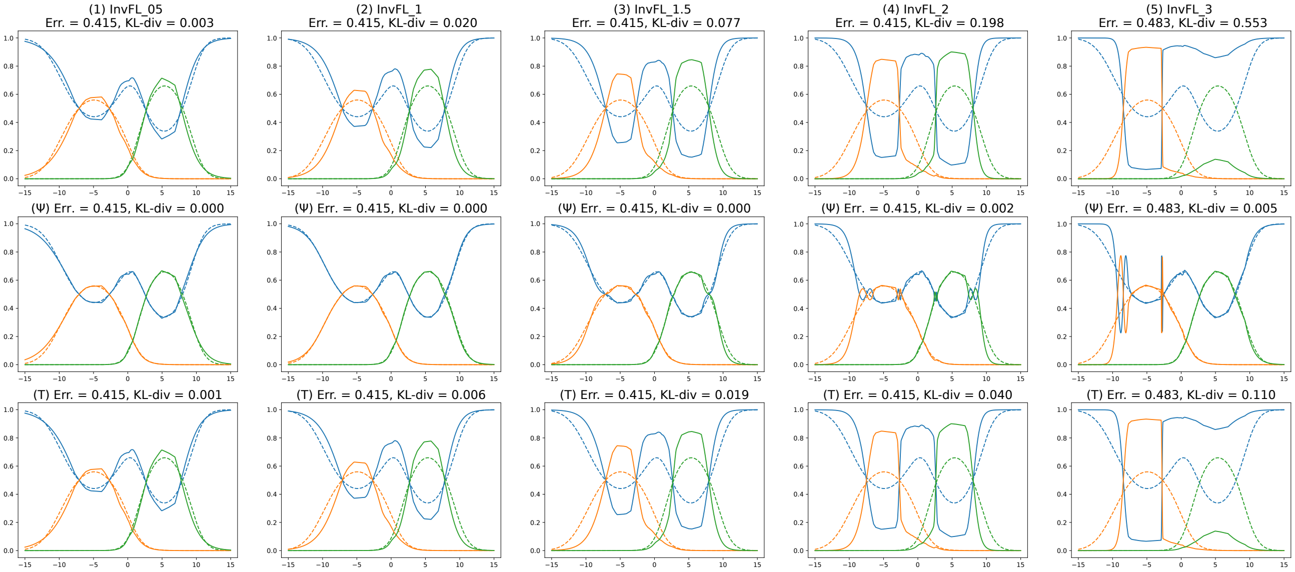


Figure 10: Numerical simulation results of the inverse focal loss, where $\gamma \in \{0.5, 1, 1.5, 2, 3\}$ from left to right, respectively. The first row indicates the performance of softmax estimation. The second row indicates the performance of $\mathbf{\Psi}$-transform. The third row indicates the performance of post-hoc temperature scaling.

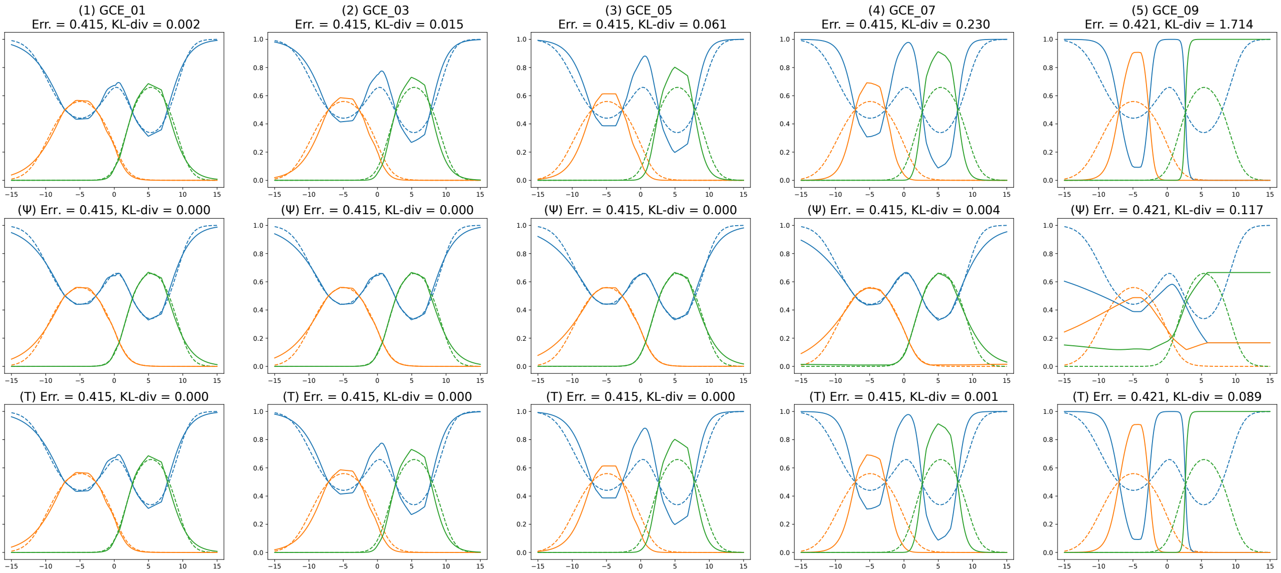


Figure 11: Numerical simulation results of the GCE loss, where $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ from left to right, respectively. The first row indicates the performance of softmax estimation. The second row indicates the performance of $\mathbf{\Psi}$-transform. The third row indicates the performance of post-hoc temperature scaling.

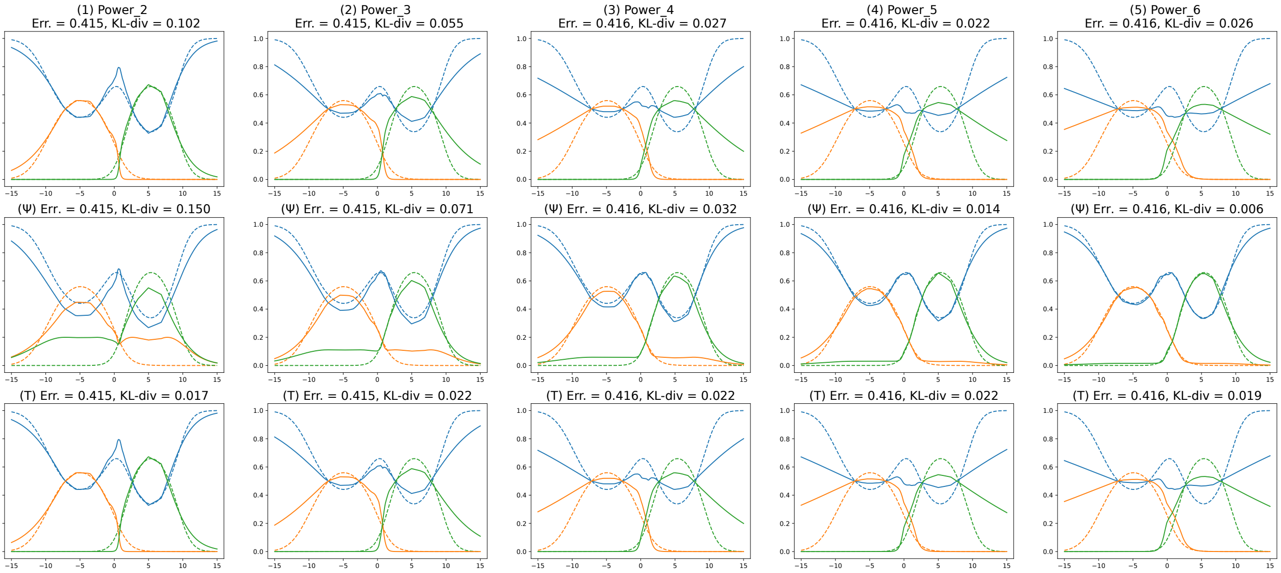
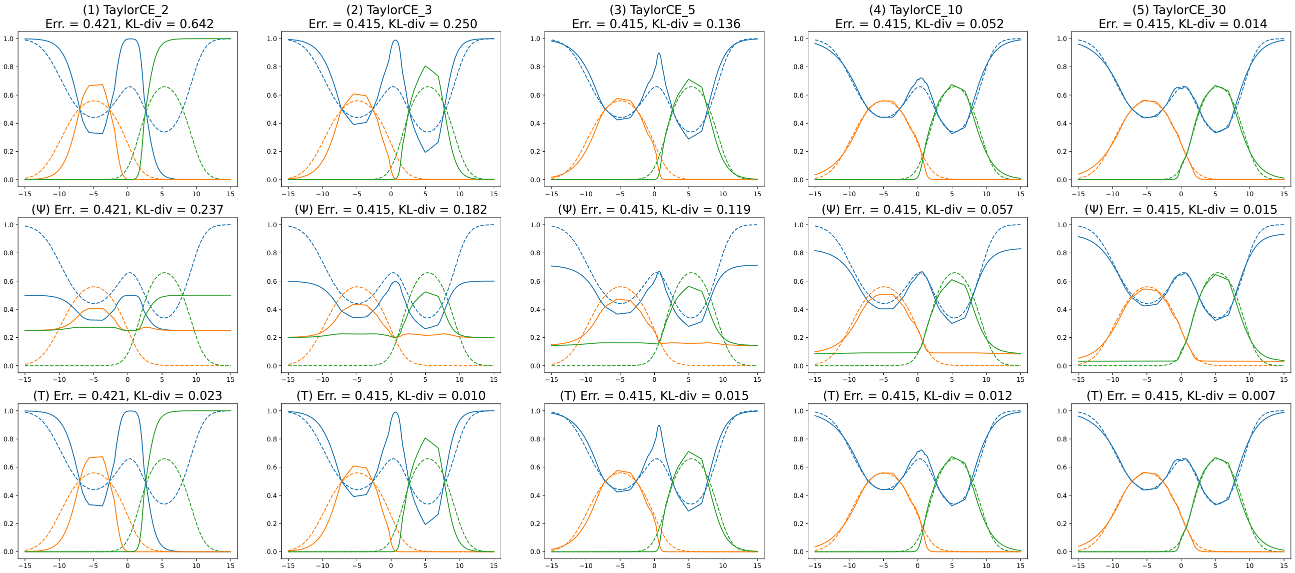


Figure 12: Numerical simulation results of the power loss, where $\gamma \in \{2, 3, 4, 5, 6\}$ from left to right, respectively. The first row indicates the performance of softmax estimation. The second row indicates the performance of $\mathbf{\Psi}$-transform. The third row indicates the performance of post-hoc temperature scaling.


(1) TaylorCE_2
Err. = 0.421, KL-div = 0.642

(2) TaylorCE_3
Err. = 0.415, KL-div = 0.250

(3) TaylorCE_5
Err. = 0.415, KL-div = 0.136

(4) TaylorCE_10
Err. = 0.415, KL-div = 0.052

(5) TaylorCE_30
Err. = 0.415, KL-div = 0.014

(Ψ) Err. = 0.421, KL-div = 0.237

(Ψ) Err. = 0.415, KL-div = 0.182

(Ψ) Err. = 0.415, KL-div = 0.119

(Ψ) Err. = 0.415, KL-div = 0.057

(Ψ) Err. = 0.415, KL-div = 0.015

(T) Err. = 0.421, KL-div = 0.023

(T) Err. = 0.415, KL-div = 0.010

(T) Err. = 0.415, KL-div = 0.015

(T) Err. = 0.415, KL-div = 0.012

(T) Err. = 0.415, KL-div = 0.007

Figure 13: Numerical simulation results of the Taylor CE loss, where $N \in \{2, 3, 5, 10, 30\}$ from left to right, respectively. The first row indicates the performance of softmax estimation. The second row indicates the performance of $\mathbf{\Psi}$-transform. The third row indicates the performance of post-hoc temperature scaling.