Functional Properties of the Focal-Entropy

Jaimin Shah
University of Minnesota,
USA

Martina Cardone
University of Minnesota,
USA

Alex Dytso
Qualcomm Flarion Technology, Inc.,
USA

Abstract

The focal-loss has become a widely used alternative to cross-entropy in class-imbalanced classification problems, particularly in computer vision. Despite its empirical success, a systematic information-theoretic study of the focal-loss remains incomplete. In this work, we adopt a distributional viewpoint and study the *focal-entropy*, a focal-loss analogue of the cross-entropy. Our analysis establishes conditions for finiteness, convexity, and continuity of the *focal-entropy*, and provides various asymptotic characterizations. We prove the existence and uniqueness of the focal-entropy minimizer, describe its structure, and show that it can depart significantly from the data distribution. In particular, we rigorously show that the focal-loss amplifies mid-range probabilities, suppresses high-probability outcomes, and, under extreme class imbalance, induces an *over-suppression regime* in which very small probabilities are further diminished. These results, which are also experimentally validated, offer a theoretical foundation for understanding the focal-loss and clarify the trade-offs that it introduces when applied to imbalanced learning tasks.

1 INTRODUCTION

In learning-based prediction and decision-making, loss functions shape the decision rule and strongly influence performance. Selecting an appropriate loss is therefore critical. A widely used choice is the *log-loss*, defined for $p \in (0,1]$ as

$$\mathsf{L}(p) = \log\left(\frac{1}{p}\right). \tag{1}$$

When predicting with a distribution Q_X, the actual outcome $x \in \mathcal{X}$ occurs with probability $P_X(x)$. The log-loss in this case is $\mathsf{L}(Q_X(x)) = \log \frac{1}{Q_X(x)}$. Averaging with respect to the true distribution P_X yields the expected log-loss, which equals the *cross-entropy* of Q_X relative to P_X. Formally, this cross-entropy over $\mathcal{X}$ is defined as

$$H(P_X, Q_X) = \begin{cases} \mathbb{E}_{X\sim P_X}\left[\mathsf{L}(Q_X(X))\right] & P_X \ll Q_X, \\ \infty & \text{otherwise.} \end{cases} \tag{2}$$

Despite its ubiquity, the log-loss often performs poorly under class imbalance – common in tasks such as object detection (Lin et al., 2017), fraud detection (He and Garcia, 2009), natural language processing (Li et al., 2020a), and various medical applications (Al Rahhal et al., 2019; Vongkulbhisal et al., 2019; Shu et al., 2019). To address this, (Lin et al., 2017) proposed the *focal-loss*, now widely adopted, especially in computer vision. Its success arises from a modification of the log-loss: a modulating pre-factor is introduced that downweights well-classified ("easy") examples and emphasizes misclassified ("hard") ones, often yielding substantial accuracy gains.

Definition 1 (Focal-loss)**.** *Given a* focus parameter $\gamma \geq 0$ *and a prediction score* $p \in (0,1]$*, the* focal-loss *is defined as*

$$\mathsf{L}_\gamma(p) = \underbrace{(1-p)^\gamma}_{\textit{focal term}} \underbrace{\log\left(\frac{1}{p}\right)}_{\textit{log-loss}}. \tag{3}$$

Although the focal-loss is empirically effective, a *comprehensive* information-theoretic treatment is still lacking. The cross-entropy has a clear foundation in information theory—minimizing it is equivalent to minimizing the Kullback–Leibler (KL) divergence between model and data—thus providing a principled

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

optimization landscape. In contrast, the focal-loss alters this landscape in ways not yet fully understood. Characterizing the geometry of the resulting objective and its minimizer(s) is therefore essential.

We adopt a distributional viewpoint and introduce the *focal-entropy*, a focal-loss analogue of the cross-entropy that compares two distributions P_X and Q_X.

Definition 2 (Focal-Entropy)**.** *Given a* focus parameter $\gamma \geq 0$*, the focal-entropy of* Q_X *relative to* P_X *over* $\mathcal{X}$ *is defined as*

$$H_\gamma(P_X, Q_X) = \begin{cases} \mathbb{E}_{X\sim P_X}[\mathsf{L}_\gamma(Q_X(X))] & P_X \ll Q_X, \\ \infty & \textit{otherwise.} \end{cases} \tag{4}$$

For $\gamma = 0$, the focal-entropy in (4) reduces to the cross-entropy in (2). A key defining property of the cross-entropy is that the minimizer over the second parameter Q_X (in the space of all probability mass functions on $\mathcal{X}$) is equal to P_X, that is,

$$\min_{Q_X} H(P_X, Q_X) = H(P_X, P_X) = H(P_X), \tag{5}$$

where the last expression is the Shannon entropy. As we show in this paper, this property, however, is no-longer true for the focal-entropy and, in general, the optimizer is not equal to P_X. In this sense, the focal-loss may not appear to be a proper loss (Brehmer and Gneiting, 2020; Buja et al., 2005). However, recent work (Bao and Charoenphakdee, 2025) has shown that the focal-loss is a variant of proper scoring rules. The structure and basic properties (e.g., existence) of the minimizer, i.e.,

$$P_\gamma^\star = \arg\min_{Q_X} H_\gamma(P_X, Q_X) \tag{6}$$

are indeed not fully understood. This paper aims to address this gap by providing a full characterization of $P_\gamma^\star$ and studying many of its properties.

1.1 Notation

The set of all probability mass functions over a (possibly countably infinite) set $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. For $P_X \in \mathcal{P}(\mathcal{X})$, its support is

$$\mathcal{S} = \{x \in \mathcal{X} : P_X(x) > 0\}. \tag{7}$$

We write $P_X \ll Q_X$ to denote absolute continuity of P_X with respect to Q_X. For P_X, we let $p_{\min} = \min_{x\in\mathcal{S}} P_X(x)$ and $p_{\max} = \max_{x\in\mathcal{S}} P_X(x)$. All logarithms are natural. The Lambert function's principal and negative branches are denoted by $\mathrm{W}_0(\cdot)$ and $\mathrm{W}_{-1}(\cdot)$, respectively (Corless et al., 1996).

1.2 Literature Review

In (Charoenphakdee et al., 2021), the authors implicitly characterized the minimizer of the focal-entropy for the finite alphabet case. They showed that the minimizing class-posterior under the focal-loss is order preserving (i.e., it maintains the same ordering as the true posterior), leading to the conclusion that the focal-loss is *classification-calibrated*. Building on this, (Mukhoti et al., 2020) demonstrated that the focal-loss not only improves accuracy on challenging datasets but also yields models that are better calibrated than those trained with the cross-entropy. Their analysis provided both theoretical justification—framing the focal-loss as a regularized KL divergence that implicitly maximizes entropy—and empirical evidence across vision and NLP tasks, including robustness under distributional shift. In this work, we give a theoretical basis for one of their key empirical findings: that focal-loss predictions exhibit higher entropy than those from log-loss. Several extensions of the focal-loss have been proposed, from modifying the focal term (Li et al., 2022, 2020b) to making the focus parameter adaptive (Ghosh et al., 2022), and even showing that focal- and log-loss arise as special cases of a broader power-series formulation (Leng et al., 2022). Related directions include class-balanced reweighting (Cui et al., 2019), margin-based methods for long-tailed learning (Cao et al., 2019), and balanced softmax corrections (Ren et al., 2020), each tackling class imbalance from complementary angles. On calibration, recent work such as dual focal-loss (Zhang et al., 2023) and entropy-based generalizations like tilted cross-entropy (Lin et al., 2021) or α-entmax losses (Peters et al., 2019) further highlight the key role of entropy-modulated objectives. Collectively, these advances situate the focal-loss within a broader family of class imbalance- and calibration-aware losses that balance sharpness, entropy, and robustness across regimes.

1.3 Paper Outline and Contributions

Section 2 examines analytical properties of the focal-loss and related functions, focusing on objects such as the inverse of its derivative, which is central to characterizing the focal-entropy minimizer.

Section 3 develops basic functional properties of the focal-entropy. Proposition 4 shows that $\gamma \mapsto H_\gamma(P_X, Q_X)$ is non-increasing and convex. Proposition 5 characterizes its asymptotic behavior for large γ, for both finite and countably infinite supports. Proposition 6 states that the focal-entropy is finite if and only if the cross-entropy is finite. Proposition 7 establishes weak lower semicontinuity and convexity of $Q \mapsto H_\gamma(P_X, Q)$. Finally, Theorem 1 proves the ex-

istence, uniqueness, and structure of the focal-entropy minimizer $P_\gamma^\star$ in (6).

Section 4 focuses on the properties of the minimizer $P_\gamma^\star$, which shed light on the ability of the focal-loss to address class imbalance. Proposition 9 shows that $P_\gamma^\star$ converges to the uniform distribution as $\gamma \to \infty$ and provides rates of convergence. To analyze how the focal-entropy transforms P_X into $P_\gamma^\star$, we study the sequence $p_{(i)} - p_{(i)}^\star$, where $p_{(i)}$ is the i-th largest entry of P_X and $p_{(i)}^\star$ is the i-th largest entry of $P_\gamma^\star$. Theorem 2 establishes that this sequence has at least one and at most two sign changes. This result enables a rigorous discussion of class imbalance, including the characterization of *exact thresholds* for when small–to–moderate probabilities are amplified (i.e., imbalance is reduced). We also show that under extreme class imbalance, i.e., when certain probabilities are very small, the focal-loss does not amplify these probabilities, but instead further suppresses them. We name this *over-suppression regime*. This finding is particularly important for practitioners, as it highlights the need to carefully select γ to avoid such regimes.

To better understand the over-suppression regime, we study trade-offs between the support size, γ, and the values of the data distribution P_X. Proposition 10 shows that for every $\gamma > 0$ and support size $|\mathcal{S}| = 2$, the over-suppression regime is not present. We also conjecture that this holds for $|\mathcal{S}| = 3$ and provide theoretical and numerical evidence supporting this conjecture. Proposition 12 and Proposition 13 establish sufficient conditions under which the over-suppression regime does not exist for arbitrary values of $|\mathcal{S}|$. Finally, Proposition 15 shows that whenever the over-suppression regime does not exist, the data distribution P_X majorizes the minimizer $P_\gamma^\star$. Taken together, these results characterize how the focal-entropy reshapes the data distribution.

Section 5 validates our results on both *synthetic* and *real* data (MNIST). Finally, Section 6 concludes the paper. Most proofs are in the appendix and the code for our experiments is available at `https://tinyurl.com/3kjmmua9`.

2 FOCAL-LOSS: ANALYTICAL PROPERTIES

We derive several analytical properties of the focal-loss and related functions that will be useful for the derivation of the results in the remaining of the paper.

In the next proposition (proof in Appendix A.1), we investigate the monotonicity and convexity of the focal-loss function in (3).

Proposition 1. *For $p \in (0,1)$ and $\gamma \geq 0$, we have*

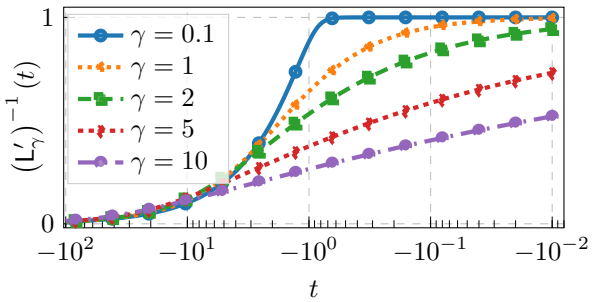

Figure 1: Inverse of L'_γ in (8).

1. *$\gamma \mapsto \mathsf{L}_\gamma(p)$ is strictly decreasing;*
2. *$p \mapsto \mathsf{L}_\gamma(p)$ is strictly decreasing with the derivative given by*
$$\mathsf{L}'_\gamma(p) = -(1-p)^{\gamma-1}\left(\gamma \log \frac{1}{p} + \frac{1-p}{p}\right) < 0; \quad (8)$$
3. *$p \mapsto \mathsf{L}_\gamma(p)$ is strictly convex with the second derivative given by*
$$\mathsf{L}''_\gamma(p) = \frac{(1-p)^{\gamma-2}}{p^2}\left[\gamma(1-\gamma)\, p^2 \log p + 2\gamma\, p(1-p) + (1-p)^2\right] > 0. \quad (9)$$

As it will be shown later, the inverse of L'_γ will play an important role in solving the first-order optimality conditions for the focal-entropy minimizer. Some of its important properties are summarized next (proof in Appendix A.2) and also illustrated in Figure 1.

Proposition 2. *For $p \in (0,1)$, the mapping $p \mapsto \mathsf{L}'_\gamma(p)$ has a well-defined inverse $\left(\mathsf{L}'_\gamma\right)^{-1} : (-\infty, 0) \to (0,1)$, which satisfies the following properties:*

1. *$\left(\mathsf{L}'_\gamma\right)^{-1}(-\infty) = 0$ and $\left(\mathsf{L}'_\gamma\right)^{-1}(0^-) = 1$;*
2. *$t \mapsto \left(\mathsf{L}'_\gamma\right)^{-1}(t)$ is strictly increasing;*
3. *$\left(\mathsf{L}'_\gamma\right)^{-1}(t) \geq 0$.*

The strict monotonicity of $\left(\mathsf{L}'_\gamma\right)^{-1}$ ensures that the mapping $t \mapsto \left(\mathsf{L}'_\gamma\right)^{-1}(t)$ is injective (see Figure 1). For $\gamma = 0$ and $\gamma = 1$, we have a closed-form expression for the inverse of (8) (see Appendix A.3).

One final function that we will require is

$$\phi_\gamma(p) = -p\, \mathsf{L}'_\gamma(p),\ p \in (0,1), \quad (10)$$

which will play an important role in Section 4, when we study properties of the minimizer of the focal-entropy. Analyzing its unimodality and bounds helps explain how the focal-loss redistributes probability masses of the data distribution, particularly in imbalanced settings. The next proposition (proof in Appendix A.4) analyzes properties of ϕ_γ in (10).

Proposition 3. *For all $\gamma \geq 0$, the mapping $p \mapsto \phi_\gamma(p)$ satisfies the following properties:*

1. $\phi_\gamma(0^+) = 1$ *and* $\phi_\gamma(1) = 0$*;*
2. $p \mapsto \phi_\gamma(p)$ *is unimodal. More specifically, the mapping $p \mapsto \phi_\gamma(p)$ is strictly decreasing if $\gamma > \kappa(p)$ and strictly increasing if $\gamma < \kappa(p)$ where*
$$\kappa(p) = \frac{1}{p} - \frac{2(1-p)}{p \log\left(\frac{1}{p}\right)}, \; p \in (0,1). \tag{11}$$
In addition,
   - *For $p \in (0,1)$, $p \mapsto \kappa(p)$ is strictly decreasing;*
   - $\kappa(p) \leq 0$ *for all* $p \in [\mathrm{w}, 1)$*, where* $\mathrm{w} = -\frac{1}{2}\mathrm{W}_0\left(-\frac{2}{\mathrm{e}^2}\right) \approx 0.2032$*;*
3. $\max_{p\in(0,1)} \phi_\gamma(p) \leq 1 + \gamma$;
4. *Let $p_\gamma^+ = \arg\max_{p\in(0,1)} \phi_\gamma(p)$, then*
$$p_\gamma^+ \leq \min\left\{\mathrm{w}, \frac{1}{\left(1+\sqrt{1+\gamma}\right)^2}\right\}; \tag{12}$$
5. *For $u \in (0, 1/3]$, it holds that $\phi_\gamma(u) \geq \phi_\gamma\left(\frac{1-u}{2}\right)$.*

3 FOCAL-ENTROPY: FUNCTIONAL PROPERTIES

In this section, we develop functional properties of the focal-entropy, proving its convexity, finiteness, and the existence and uniqueness of its minimizer.

3.1 Focal-Entropy vs. γ

We here derive a few properties that highlight how the focal-entropy in (4) varies as a function of γ. The first property is directly derived from the inspection of (4) (see also Appendix B.1, where some additional properties of the mapping $\gamma \mapsto H_\gamma(P_X, Q_X)$ are provided).

Proposition 4. *Fix some $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ such that $P_X \ll Q_X$. For $\gamma \geq 0$, the mapping $\gamma \mapsto H_\gamma(P_X, Q_X)$ is non-increasing and convex. Thus, we have that*
$$0 \leq H_\gamma(P_X, Q_X) \leq H_0(P_X, Q_X) = H(P_X, Q_X). \tag{13}$$

The next proposition (proof in Appendix B.2) presents a limit for the focal-entropy when the focus parameter γ tends to infinity.

Proposition 5. *Fix some $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ such that $P_X \ll Q_X$ and $H(P_X, Q_X) < \infty$. Then, it holds that*
$$\lim_{\gamma\to\infty} H_\gamma^{\frac{1}{\gamma}}(P_X, Q_X) = \begin{cases} \max_{x\in\mathcal{S}} \bar{Q}_X(x) & |\mathcal{S}| < \infty, \\ 1 & |\mathcal{S}| = \infty, \end{cases} \tag{14}$$
where $\bar{Q}_X(x) = 1 - Q_X(x)$ and $\mathcal{S}$ is defined in (7).

We highlight that, when the support size $|\mathcal{S}|$ is finite, minimizing the right-hand side of (14) over Q_X yields the uniform distribution over $\mathcal{S}$. A more precise statement of uniformity as $\gamma \to \infty$ is given in Proposition 9.

3.2 Basic Functional Properties

We derive some basic functional properties of the focal-entropy in (4). We start with the next proposition (proof in Appendix B.3), which gives a sufficient and necessary condition for the focal-entropy to be finite.

Proposition 6. *Fix some $\gamma \geq 0$. Then, $H_\gamma(P_X, Q_X) < \infty$ if and only if $H(P_X, Q_X) < \infty$.*

For the remainder of the paper, we consider only distributions with *finite* focal-entropy; see (Cover and Thomas, 2006) for an example of infinite entropy. The next proposition (proof in Appendix B.4) presents additional functional properties of the focal-entropy.

Proposition 7. *For any $\gamma \geq 0$ and fixed P_X, it holds that the mapping $Q \mapsto H_\gamma(P_X, Q)$:*

1. *is weakly lower semicontinuous*[1]*; and*
2. *is convex. Moreover, it is* strictly *convex over the probability space defined on $\mathcal{S}$ in* (7)*.*

Remark 1. Proposition 7 shows that the focal-entropy is weakly lower semicontinuous. We note that, in general, the focal-entropy is not weakly continuous (see Appendix B.5 for an example).

In (Mukhoti et al., 2020), the authors provided an inequality between the focal-entropy and cross-entropy, which allows certain learning conclusions to be made. In Appendix B.6, we provide a novel relationship that bounds the difference between their infimum values.

3.3 Optimization of the Focal-Entropy

The minimum of the cross-entropy in (2) (see (5)) follows from the non-negativity of the relative entropy, or equivalently Gibbs' inequality. We now show a similar result for the focal-entropy.

Theorem 1. *Fix some $P_X \in \mathcal{P}(\mathcal{X})$ with support $\mathcal{S}$ as in (7) and $\gamma \geq 0$. Then, there exists a unique $0 < \alpha_\gamma^\star < \infty$ such that*
$$P_\gamma^\star(x) = \begin{cases} (\mathsf{L}_\gamma')^{-1}\left(-\frac{\alpha_\gamma^\star}{P_X(x)}\right) & x \in \mathcal{S}, \\ 0 & x \notin \mathcal{S}, \end{cases} \tag{15}$$

[1] A functional $F : \mathcal{P}(\mathcal{X}) \to (-\infty, \infty]$ is said to be *weakly lower semicontinuous* if for every sequence $(Q_n)_{n\geq1}$ such that $Q_n \to Q$ (weak convergence), we have that $F(Q) \leq \liminf_{n\to\infty} F(Q_n)$.

is a valid distribution. Moreover, for all $Q_X \in \mathcal{P}(\mathcal{X})$, it holds that

$$H_\gamma(P_X, Q_X) \geq H_\gamma(P_X, P_\gamma^\star) + \alpha_\gamma^\star \left(1 - \sum_{x\in\mathcal{S}} Q_X(x)\right) \quad (16)$$

$$\geq H_\gamma(P_X, P_\gamma^\star), \quad (17)$$

and $P_\gamma^\star = \arg\min_{Q_X} H_\gamma(P_X, Q_X)$ is the unique minimizer for the focal-entropy.

Proof. For $P_\gamma^\star$ in (15) to be a valid distribution, the constant $\alpha_\gamma^\star$ must be a positive root of the equation $F(\alpha_\gamma^\star) = 1$, where

$$F(\alpha) = \sum_{x\in\mathcal{S}} (\mathsf{L}_\gamma')^{-1}\left(-\frac{\alpha}{P_X(x)}\right), \quad \alpha > 0. \quad (18)$$

From Proposition 2, each summand in (18) is continuous and strictly decreasing in $\alpha > 0$ having values on $(0, 1)$. Thus, $\alpha \mapsto F(\alpha)$ is continuous and strictly decreasing with $F(0^+) = |\mathcal{S}|$ and $F(+\infty) = 0$ (see Proposition 2). Then, by the intermediate value theorem, there is a unique positive root $\alpha_\gamma^\star$, yielding a valid $P_\gamma^\star$.

To show the second part of the theorem, recall from Proposition 1 that $p \mapsto \mathsf{L}_\gamma(p)$ is strictly convex and so it can be lower bounded by a tangent line as follows,

$$\mathsf{L}_\gamma(Q(X)) \quad (19)$$

$$\geq \mathsf{L}_\gamma\left(P_\gamma^\star(X)\right) + \mathsf{L}_\gamma'\left(P_\gamma^\star(X)\right)\left(Q(X) - P_\gamma^\star(X)\right) \quad (20)$$

$$= \mathsf{L}_\gamma\left(P_\gamma^\star(X)\right) - \frac{\alpha_\gamma^\star}{P_X(X)}\left(Q(X) - P_\gamma^\star(X)\right). \quad (21)$$

Consequently, for any choice of Q, we arrive at

$$H_\gamma(P_X, Q) \quad (22)$$

$$= \mathbb{E}\left[\mathsf{L}_\gamma(Q(X))\right] \quad (23)$$

$$\geq \mathbb{E}[\mathsf{L}_\gamma\left(P_\gamma^\star(X)\right)] - \mathbb{E}\left[\frac{\alpha_\gamma^\star}{P_X(X)}\left(Q(X) - P_\gamma^\star(X)\right)\right] \quad (24)$$

$$= \mathbb{E}[\mathsf{L}_\gamma\left(P_\gamma^\star(X)\right)] - \alpha_\gamma^\star\left(\sum_{x\in\mathcal{S}} Q(x) - 1\right) \quad (25)$$

$$\geq \mathbb{E}[\mathsf{L}_\gamma\left(P_\gamma^\star(X)\right)] \quad (26)$$

$$= H_\gamma(P_X, P_\gamma^\star), \quad (27)$$

where all the expectations are taken with respect to $X \sim P_X$. The fact that $P_\gamma^\star$ in (15) is the unique minimizer for the focal-entropy is due to the convexity result in Proposition 7. □

Figure 2 illustrates how a distribution P_X with $|\mathcal{S}| = 3$ is transformed into $P_\gamma^\star$ in (15) for different values of γ. The smallest probability (at $x = x_1$) is amplified, while the largest (at $x = x_2$) is reduced. Intuitively, the transformation shifts the mass from high- to low-probability outcomes, mitigating imbalance. Section 4 gives a complete characterization of which probabilities are amplified or suppressed.

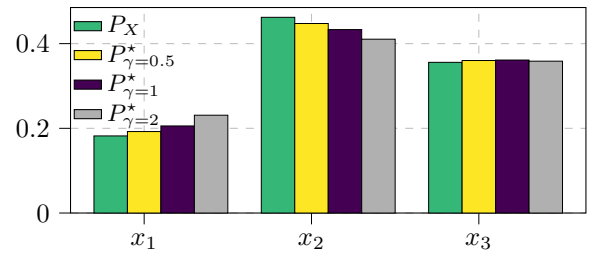

Figure 2: Example of how a P_X with $|\mathcal{S}| = 3$ is transformed into $P_\gamma^\star$ in (15) for different values of γ.

Some immediate consequences of Theorem 1 (with Proposition 2) are provided in the next corollary (proof in Appendix B.8).

Corollary 1. *The distribution $P_\gamma^\star$ in Theorem 1 satisfies the following properties:*

1. *$P_\gamma^\star \ll P_X$;*
2. *For any $x_1, x_2 \in \mathcal{S}$*
$$P_X(x_1) \geq P_X(x_2) \implies P_\gamma^\star(x_1) \geq P_\gamma^\star(x_2); \quad (28)$$
3. *$P_\gamma^\star = P_X$ if and only if $\gamma = 0$ or P_X is a uniform distribution.*

Remark 2. It is useful to think of the mapping $P_X \mapsto P_\gamma^\star$ as an operator. The inverse of this operator is well-defined and has a closed-form expression given by

$$P_X(x) = \frac{\beta}{\mathsf{L}_\gamma'(P_\gamma^\star(x))}, \; x \in \mathcal{S}, \quad (29)$$

$$\beta = \left(\sum_{v\in\mathcal{S}} \frac{1}{\mathsf{L}_\gamma'(P_\gamma^\star(v))}\right)^{-1}. \quad (30)$$

In some cases, this inverse operator is easier to handle due to its closed form. This view relates to (Charoenphakdee et al., 2021, Theorem 11), which derived (29) for finite alphabets via Lagrangian duality. In contrast, Theorem 1 extends the result to countably infinite alphabets and gives a more geometric proof based on the inverse of the focal-loss derivative. Our setting is also more natural in practice: since the optimizer $P_\gamma^\star$ is usually unknown, one typically works with the forward mapping $P_X \mapsto P_\gamma^\star$.

The minimizer in (15) depends on $\alpha_\gamma^\star$ and hence, finding bounds on $\alpha_\gamma^\star$ helps to reduce the search space. The next proposition (proof in Appendix B.9) provides bounds on $\alpha_\gamma^\star$, which will be leveraged for theoretical derivations in the remainder of the paper.

Proposition 8. *For all $\gamma \geq 0$, it holds that*

$$\alpha_\gamma^\star \leq \max_{t\in(0,1)} \phi_\gamma(t) \leq 1 + \gamma. \quad (31)$$

Moreover, suppose that $|\mathcal{S}| = N < \infty$. Then:

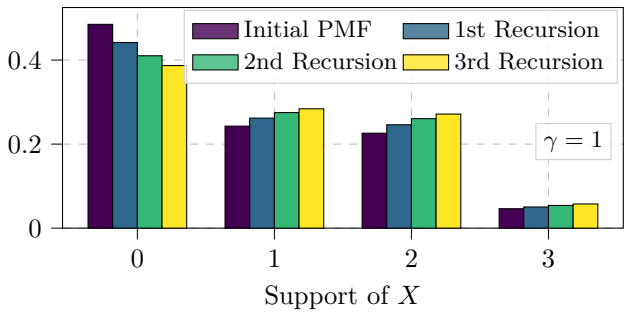

Figure 3: Initial probability mass function in (35).

- *For all $\gamma \geq 0$, it holds that*

$$p_{\min}\, c_{N,\gamma} \leq \alpha^\star_\gamma \leq p_{\max}\, c_{N,\gamma}, \tag{32}$$

where $c_{N,\gamma} = -\mathsf{L}'_\gamma(1/N) > 0$.

- *For all $\gamma > \kappa(p_{\min})$, with $\kappa(\cdot)$ being defined in (11), it holds that*

$$\phi_\gamma(p_{\max}) \leq \alpha^\star_\gamma \leq \phi_\gamma(p_{\min}). \tag{33}$$

- *For all $\gamma < \kappa(p_{\max})$, with $\kappa(\cdot)$ being defined in (11), it holds that*

$$\phi_\gamma(p_{\min}) < \alpha^\star_\gamma < \phi_\gamma(p_{\max}). \tag{34}$$

Remark 3. The lower and upper bounds in (32), (33), and (34) match when P_X is a uniform distribution.

Remark 4. It is well known that the minimizer of the cross-entropy is *idempotent*: if the minimization procedure is initialized at a distribution $P_0^\star$, the optimizer coincides with the initialization and no further change occurs. As illustrated in Figure 3, the optimizer $P^\star_\gamma$ in (15) is, in general, *not idempotent*. The results shown in Figure 3 were obtained starting from the initial probability mass function

$$P_X = \begin{bmatrix} 0.485 & 0.243 & 0.226 & 0.046 \end{bmatrix}. \tag{35}$$

We then numerically determined (using the bounds in Proposition 8) the value of $\alpha^\star_\gamma$ satisfying the normalization condition $F(\alpha^\star_\gamma) = 1$, where $F(\cdot)$ is defined in (18). Using this value, we computed the corresponding optimizer $P^\star_\gamma$ via (15) (Recursion 1 in Figure 3), which differs from the initial distribution P_X. Repeating this procedure by reapplying the optimization to the newly obtained distribution (Recursions 2 and 3 in Figure 3) reveals that successive iterates continue to change. Hence, unlike the cross-entropy case, focal-entropy minimization induces a non-trivial transformation of the distribution under repeated application.

4 PROPERTIES OF THE OPTIMIZER

In this section, we explore properties of the optimizer $P^\star_\gamma$ in (15). These are critical since they can shed light on important characteristics of the focal-loss, such as how it helps with class imbalance.

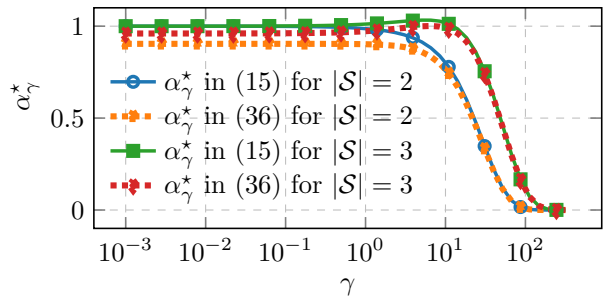

Figure 4: Comparison of $\alpha^\star_\gamma$ in (15) with its approximation in (36).

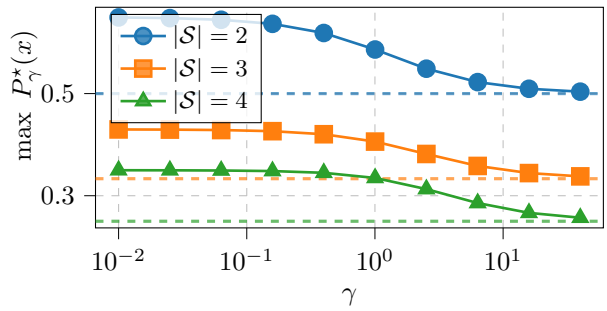

Figure 5: Convergence of $P^\star_\gamma$ to $1/|\mathcal{S}|$.

4.1 Large γ Asymptotics of $P^\star_\gamma$

Our goal here is to understand how $P^\star_\gamma$ in (15) is affected as the value of γ increases. The next proposition (proof in Appendix C.1) answers this question.

Proposition 9. *Suppose that $|\mathcal{S}| < \infty$. Then, as $\gamma \to \infty$, we have that*

$$\alpha^\star_\gamma = -\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\mathsf{HM}(P_X) + O\left(\gamma\left(1-\frac{1}{|\mathcal{S}|}\right)^\gamma\right), \tag{36}$$

$$P^\star_\gamma(x) = \frac{1}{|\mathcal{S}|} + O\left(\frac{1}{\gamma}\right), \tag{37}$$

where $\mathsf{HM}(P_X) = \frac{|\mathcal{S}|}{\sum_{x\in\mathcal{S}} \frac{1}{P_X(x)}}$ is the harmonic mean *of P_X.*

Figure 4 compares $\alpha^\star_\gamma$ in (15) with its approximation in (36) for the initial distributions reported in Appendix C.1, and shows that the approximation is highly accurate. Similarly, Figure 5 shows the convergence of $P^\star_\gamma$ to the uniform distribution for different values of $|\mathcal{S}|$, again using the same initial distributions.

4.2 The Three Bins Property

We wish to understand when $P^\star_\gamma$ 'dominates' P_X and vice versa. To this end, we define the sequence:

$$d_i = p_{(i)} - p^\star_{(i)},\ i \in \mathbb{N}, \tag{38}$$

where $p_{(i)}$ is the i-th largest entry of P_X and $p^\star_{(i)}$ is the i-th largest entry of $P^\star_\gamma$ in (15). By Corollary 1, P_X and $P^\star_\gamma$ share the same ordering; thus, d_i is well defined, preserving the correspondence between ordered entries. An immediate consequence of this and the fact that probabilities sum to one is the following lemma.

Lemma 1. *If P_X is uniform or $\gamma = 0$, then $d_i = 0$ for all $i \in \mathbb{N}$. Otherwise, the sequence $\{d_i\}$ has* at least *one sign change.*

Before providing the main result of this section, which can be thought of as a converse to Lemma 1, we present the following preparatory lemma.

Lemma 2. *For any fixed $\gamma > 0$, $\phi_\gamma(p) - \alpha^\star_\gamma = 0$ has at least one and at most two roots. In particular,*

- *$p_{\gamma,a} \in [0, p^+_\gamma)$ denotes the smallest root, which exists only if $\alpha^\star_\gamma \geq 1$. For $\alpha^\star_\gamma < 1$, we set $p_{\gamma,a} = 0$.*
- *$p_{\gamma,b} \in (p^+_\gamma, 1)$ denotes the largest root, which always exists.*

Proof. We recall that $\phi_\gamma(p)$ is unimodal (Proposition 3); hence $\phi_\gamma(p) - \alpha^\star_\gamma = 0$ has at most two roots. At least one root exists since

$$\alpha^\star_\gamma \leq \max_{p\in(0,1)} \phi_\gamma(p), \tag{39}$$

as shown in Proposition 8. Moreover, since $\phi_\gamma(0^+) = 1$, $\phi_\gamma(1) = 0$, and ϕ_γ is unimodal, if $\alpha^\star_\gamma < 1$ there is only one root, $p_{\gamma,b}$. This completes the proof. □

The next result (proof in Appendix C.2) characterizes how the sequence of sign changes in $\{d_i\}$ behaves.

Theorem 2. *Suppose that P_X is* not *uniform and $\gamma > 0$. Then, the following holds:*

- *if $p_{(i)} \in (0, p_{\gamma,a}]$, then $d_i \geq 0$;*
- *if $p_{(i)} \in (p_{\gamma,a}, p_{\gamma,b})$, then $d_i < 0$;*
- *if $p_{(i)} \in [p_{\gamma,b}, 1)$, then $d_i \geq 0$;*

where $0 \leq p_{\gamma,a} \leq p^+_\gamma \leq p_{\gamma,b} < 1$ are defined in Lemma 2.

Figure 6 illustrates the result in Theorem 2. In words, Theorem 2 has the following consequences:

- **Small-to-moderate probabilities ($p_{\gamma,a}$, $p_{\gamma,b}$).** In this regime $d_i < 0$ and hence, $p^\star_{(i)} > p_{(i)}$. The focal-loss amplifies these probabilities, which is the principal mechanism by which it mitigates class imbalance, i.e., the probability mass is shifted toward mid–range entries.

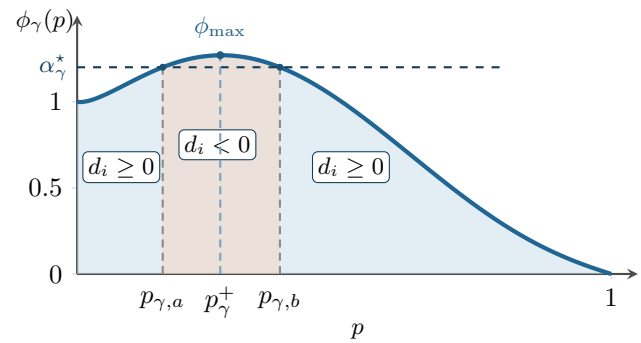

Figure 6: Illustration of Theorem 2.

- **Large probabilities ($p_{\gamma,b}$, 1).** Here $d_i \geq 0$ and hence, $p^\star_{(i)} \leq p_{(i)}$. The focal-loss downweights high-probability ("easy") samples. Part of this effect is normalization: since mid-range probabilities are amplified and since $\sum_i p^\star_{(i)} = 1$, then the largest entries must decrease.
- **Very small probabilities (0, $p_{\gamma,a}$].** In this regime, which we refer to as *over-suppression regime*, we also have $d_i \geq 0$ and hence, $p^\star_{(i)} \leq p_{(i)}$. Thus, extremely small probabilities are further suppressed, which can, in principle, exacerbate imbalance at the extreme tail. However, as shown in (12), for sufficiently large values of γ the interval $(0, p_{\gamma,a}]$ is empty or negligible. Thus, while in practice this effect is typically muted, one should not ignore this regime especially in extremely imbalanced regimes.

An immediate consequence of Theorem 2 is provided in the next corollary (see also Figure 6).

Corollary 2. *Suppose that P_X is* not *uniform and $\gamma > 0$. Then, the sequence $\{d_i\}, i \in \mathbb{N}$, has at least one and at most two sign changes.*

4.3 Sign Changes for Small Support Sizes

We show that when the support is binary (proof in Appendix C.3) and possibly ternary, we have $p_{\min} > p_{\gamma,a}$, i.e., the *over-suppression regime* never occurs.

Proposition 10. *Let $|\mathcal{S}| = 2$. Then, for all $\gamma > 0$, it holds that $d_1 \geq 0$ and $d_2 < 0$.*

Conjecture 1. *Let $|\mathcal{S}| = 3$. Then, for all $\gamma > 0$, it holds that $p_{\min} > p_{\gamma,a}$, i.e., $d_3 < 0$ (i.e., the over-suppression regime is absent for $|\mathcal{S}| = 3$).*

We now provide numerical and theoretical evidence in support of Conjecture 1. Numerically, we carried out extensive simulations (see Figure 7) and we could not find cases where $d_3 \geq 0$. To provide theoretical evidence, we use the next result (proof in Appendix C.4).

Proposition 11. *Let $|\mathcal{S}| = 3$. Then, for all $\gamma > 0$, it holds that $d_1 \geq 0$.*

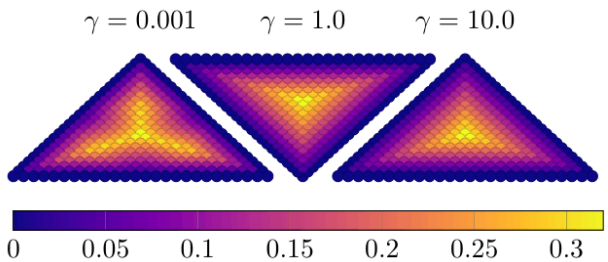

Figure 7: Heatmap of $p_{\min} - p_{\gamma,a}$ over the ternary probability simplex.

Now, since $d_1 \geq 0$, two cases can occur:

1. $d_2 \geq 0$ (i.e., $p_{(2)} > p_{\gamma,b}$): in this case, from Corollary 2, we need $d_3 < 0$, i.e., $p_{\min} > p_{\gamma,a}$. Therefore, for this case, Conjecture 1 holds.
2. $d_2 < 0$ (i.e., $p_{\gamma,a} < p_{(2)} < p_{\gamma,b}$): For this regime, proving that $d_3 < 0$ (i.e., $p_{\min} > p_{\gamma,a}$) is an open problem. In other words, this is the only regime that must be addressed to prove Conjecture 1.

Remark 5. Although for the case $|\mathcal{S}| \leq 3$ it appears that the *over-suppression regime* never occurs for all $\gamma > 0$, the same is not true for $|\mathcal{S}| \geq 4$. Appendix C.5 gives an example with $|\mathcal{S}| = 4$ where $p_{\min} < p_{\gamma,a}$.

4.4 Interplay between γ, Cardinality of the Support, and P_X

We now seek to study the interplay between γ, the cardinality of the support $\mathcal{S}$, and the values of P_X. In particular, we provide sufficient conditions under which the over-suppression regime does not exist.

Proposition 12. *Suppose that $|\mathcal{S}| < \infty$ and*

$$-p_{\max}\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) < 1. \tag{40}$$

Then, $p_{\gamma,a} = 0$, that is, the over-suppression regime does not exist.

Proof. Combining the upper bound on $\alpha^\star_\gamma$ in (32) with the assumption in (40), we arrive at $\alpha^\star_\gamma < 1$. By Lemma 2, $\alpha^\star_\gamma < 1$ implies that $p_{\gamma,a} = 0$. This concludes the proof of Proposition 12. □

The condition in (40) has three degrees of freedom, namely: $p_{\max}$, γ, and $|\mathcal{S}|$. This condition can be further loosened to

$$-\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) < 1. \tag{41}$$

From Figure 8, for each $|\mathcal{S}|$, there exists $\bar{\gamma}$, that depends only on $|\mathcal{S}|$, such that $p_{\gamma,a} = 0$ for all $\gamma > \bar{\gamma}$, i.e., eventually the over-suppression regime vanishes.

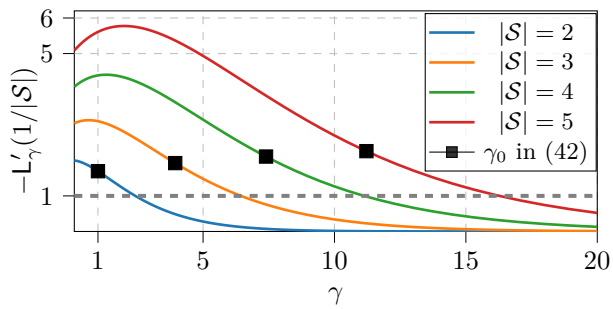

Figure 8: Illustration of conditions in (41) and (42).

The next proposition (proof in Appendix C.6) provides an additional bound on the minimum value of γ that ensures that the over-suppression regime cannot occur.

Proposition 13. *Let $\kappa(\cdot)$ be given in* (11). *If $\gamma > \kappa(p_{\min})$, then $p_{\min} > p_{\gamma_a}$, that is, the over-suppression regime does not exist.*

We now present a final result giving a sufficient condition for $p_{\max} > p_{\gamma,b}$, i.e., $d_1 \geq 0$. In other words, the next proposition (proof in Appendix C.7) ensures that the maximum probability is reduced.

Proposition 14. *Suppose that $|\mathcal{S}| < \infty$ and $\gamma > \gamma_0$ where*

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_{-1}\left(-n\exp(-n)\right), \tag{42}$$

with $n = \log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)$. Then, $p_{\max} > p_{\gamma,b}$, i.e., $d_1 > 0$.

Figure 8 also reports γ_0 from (42) for different values of $|\mathcal{S}|$. The condition $\gamma > \gamma_0$ in Proposition 14 is less stringent than (41), since the latter ensures both $d_1 \geq 0$ and $p_{\gamma,a} = 0$, while the former guarantees only $d_1 \geq 0$.

4.5 Majorization

It has been observed empirically (Mukhoti et al., 2020) that the focal-entropy optimizer $P^\star_\gamma$ typically has higher entropy than the original P_X. We now formalize this observation for distributions with finite support size. We start with the following definition.

Definition 3 (Majorization)**.** *For two vectors $x, y \in \mathbb{R}^n$, we say that x majorizes y (notation $x \succ y$) when*

$$\sum_{i=1}^{k} x_{(i)} \geq \sum_{i=1}^{k} y_{(i)}, \text{ for all } k = 1, \ldots, n, \tag{43}$$

where $x_{(i)}$ is the i-th largest entry of x.

An immediate consequence of Theorem 2 is given in the next proposition (proof in Appendix C.8).

Proposition 15. *Suppose that $p_{\min} > p_{\gamma_a}$. Then, P_X majorizes $P^\star_\gamma$, that is, $P_X \succ P^\star_\gamma$.*

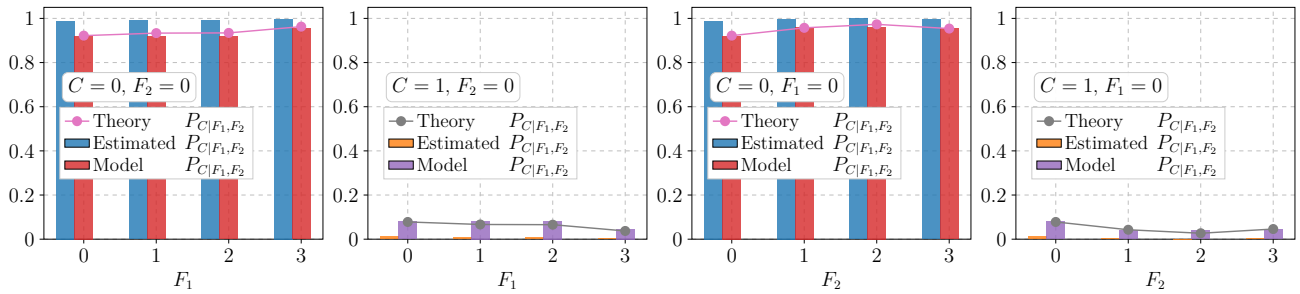

Figure 9: MNIST dataset: For each class, we fixed one of the features at its modal value. "Model $P_{C|F_1,F_2}$" is the predicted probability of class C for each of the 16 possible bin combinations (i.e., four bins for F_1 and four bins for F_2), which is estimated by using a neural network trained by using the focal-loss in (3) with $\gamma = 1$; "Estimated $P_{C|F_1,F_2}$" is the empirical approximation of the posterior (since the true posterior is unknown), which is obtained by computing, for each of the 16 bin combinations, the proportion of samples with $C = 0$ and $C = 1$; and "Theory $P_{C|F_1,F_2}$" is the $P^\star_\gamma$ in (15) obtained by using "Estimated $P_{C|F_1,F_2}$" as input.

Note that sufficient conditions for which $p_{\min} > p_{\gamma_a}$ are provided in Proposition 12 and Proposition 13. Combining the result in Proposition 15 with Schur-convexity, we have the following consequences.

Corollary 3. *Suppose that $p_{\min} > p_{\gamma_a}$. Then:*

- Entropy increase*:*

$$H(P^\star_\gamma) \geq H(P_X), \tag{44}$$

which follows since the Shannon entropy is Schur-concave (Kvålseth, 2022). The above result generalizes to the Rényi entropy (Kvålseth, 2022).

- Non-uniformity decrease*: By subtracting* $\log(|\mathcal{S}|)$ *from both sides of* (44)*, we have that*

$$D_{\mathsf{KL}}(P_X \| P_U) \geq D_{\mathsf{KL}}(P^\star_\gamma \| P_U), \tag{45}$$

where P_U is the uniform distribution.

Remark 6. Corollary 3 formalizes the empirical finding of (Mukhoti et al., 2020) that $H(P^\star_\gamma) \geq H(P_X)$. This is significant, since higher-entropy predictions help reduce overconfidence (Mukhoti et al., 2020).

5 EXPERIMENTAL VALIDATION

In this section, we validate our theoretical results on both *synthetic* and *real* data. For the synthetic case (Appendix D), we assume knowledge of the joint distribution over class labels and features, enabling direct theoretical validation.

For the *real* data, we used MNIST in binary classification: $C = 0$ for 'digit is not 1' and $C = 1$ for 'digit is 1', yielding class imbalance (53,258 of 60,000 images in class 0). From each 28×28 image, we extracted two zoning-style features capturing spatial asymmetry (Impedovo and Pirlo, 2014): F_1, the intensity ratio of upper vs. lower halves, and F_2, the intensity ratio of left vs. right halves. Each image was represented by these two features, which were then quantized into four bins (empirical quantiles (Dougherty et al., 1995)).

We trained a neural network (same architecture as in Appendix D) using the focal loss in (3) with $\gamma = 1$. The trained model estimated the probability of C for each of the 16 bin combinations (four bins for F_1, four for F_2)—see the 'Model $P_{C|F_1,F_2}$' bars in Figure 9.

Since the true posterior is unknown, we approximated it by computing the empirical class proportions within each bin combination ('Estimated $P_{C|F_1,F_2}$' in Figure 9). Using this estimated posterior in (15), we derived $P^\star_\gamma$ for $\gamma = 1$ ('Theory $P_{C|F_1,F_2}$'). Figure 9 shows that the network outputs closely match $P^\star_\gamma$ (maximum difference 0.017), supporting our theory and indicating convergence to the global minimum under focal loss.

6 CONCLUSION

We conducted a systematic study of the focal-entropy. We showed that it preserves desirable features (finiteness and convexity), while exhibiting behaviors that distinguish it fundamentally from the cross-entropy. We established the existence and uniqueness of its minimizer and described how it reshapes distributions by amplifying mid-range probabilities, suppressing large probability outcomes, and, under certain conditions, over-suppressing very small probabilities. This finding is particularly important for practitioners, as it highlights the need to carefully select γ to avoid the *over-suppression* regime. These results clarify the trade-offs that the focal-loss introduces and provide a theoretical foundation for its use in imbalanced learning. We hope that this perspective will guide further theoretical developments and support more principled practical applications of the focal-loss. An interesting future direction is to study focal-loss behavior with soft, real-valued labels (Jeong et al., 2023) in classification task.

References

M. M. Al Rahhal, Y. Bazi, H. Almubarak, N. Alajlan, and M. Al Zuair. Dense convolutional networks with focal loss and image generation for electrocardiogram classification. *IEEE Access*, 7:182225–182237, 2019.

Han Bao and Nontawat Charoenphakdee. Calm composite losses: Being improper yet proper composite. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Jonas R Brehmer and Tilmann Gneiting. Properization: Constructing proper scoring rules via Bayes acts. *Annals of the Institute of Statistical Mathematics*, 72(3):659–673, 2020.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 3:13, 2005.

K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 32, pages 1565–1576, 2019.

N. Charoenphakdee, J. Vongkulbhisal, N. Chairatanakul, and M. Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2021.

R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambert w function. *Advances in Computational Mathematics*, 5 (1):329–359, 1996.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (2nd edn)*. Wiley, 2006.

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277, 2019.

J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proc. 1995*, pages 194–202. Elsevier, 1995.

A. Dytso and M. Cardone. Lossy source coding with focal loss. *arXiv preprint arXiv:2504.19913*, 2025.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, New York, 2 edition, 1999.

A. Ghosh, T. Schaaf, and M. Gormley. Adafocal: Calibration-aware adaptive focal loss. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 35:1583–1595, 2022.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

D. Impedovo and G. Pirlo. Zoning methods for handwritten character recognition: A survey. *Pattern Recognit.*, 47(3):969–981, 2014.

M Jeong, M. Cardone, and A. Dytso. Demystifying the optimal performance of multi-class classification. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=p9k5MS0JAL`.

T. O. Kvålseth. Entropies and their concavity and schur-concavity conditions. *IEEE Access*, 10:96006–96015, 2022.

Z. Leng, M. Tan, C. Liu, E. D. Cubuk, X. Shi, S. Cheng, and D. Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. *arXiv preprint arXiv:2204.12511*, 2022.

X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4669, 2020a.

X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 33:21002–21012, 2020b.

X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang. Generalized focal loss: Towards efficient representation learning for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3139–3153, 2022.

J. Lin, X. Jin, and C. Zhang. Tce: Tilted cross-entropy loss for classification. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4061–4070, 2021.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 2980–2988, 2017.

J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 15288–15299, 2020.

B. Peters, V. Niculae, and A. F. T. Martins. Sparse sequence-to-sequence models. In *Proc. 57th Annu. Meeting Assoc. Computational Linguistics (ACL)*, pages 1504–1519, 2019.

M. Ren, W. Yu, J. Ma, S. Zhao, H. Li, and R. Urtasun. Balanced meta-softmax for long-tailed visual recog-

nition. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, pages 4175–4186, 2020.

W. Shu, S. Wang, Q. Chen, Y. Hu, Z. Cai, and R. Lin. Pathological image classification of breast cancer based on residual network and focal loss. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pages 211–214, 2019.

J. Vongkulbhisal, P. Vinayavekhin, and M. Visentini-Scarzanella. Unifying heterogeneous classifiers with distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3175–3184, 2019.

Y. Zhang, C. Chen, H. Shi, M. Song, Y. Wu, and J. Zhang. Dual focal loss for calibrated classification. In *Proc. 40th Int. Conf. Machine Learning (ICML)*, 2023.

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. Please, see Section 5 and Appendix D.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Not Applicable.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Not Applicable.**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes.**
   (b) Complete proofs of all theoretical results. **Yes. The majority of the proofs can be found in the supplementary material.**
   (c) Clear explanations of any assumptions. **Yes.**

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 code used to generate our experimental results can be found at `https://tinyurl.com/3kjmmua9`.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes. Please, see Section 5 and Appendix D.**
   (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). **Not Applicable.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Not Applicable.**

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 for Section 2

A.1 Proof of Proposition 1

To show the first property, note that for $p \in (0,1)$, the function $\gamma \mapsto p^{(1-p)^\gamma}$ is strictly increasing and hence, $\gamma \mapsto \mathsf{L}_\gamma(p)$ is strictly decreasing.

The first and second derivatives can be obtained through a straightforward calculation. It is also trivial to see that $\mathsf{L}'_\gamma(p) < 0$ for all $p \in (0,1)$ and $\gamma \geq 0$.

We now show that $\mathsf{L}''_\gamma(p) > 0$ for all $p \in (0,1)$ and $\gamma \geq 0$, which implies that $p \mapsto \mathsf{L}_\gamma(p)$ is strictly convex on $(0,1)$. We consider two cases separately:

- **Case $\gamma > 1$ or $\gamma = 0$:** it is trivial to see that the right-hand side of (9) is always positive;
- **Case** $0 < \gamma \leq 1$: by using the fact that $\log(p) \geq 1 - \frac{1}{p}$ for all $p > 0$, we can lower bound the second order derivative in (9) as follows,

$$\mathsf{L}''_\gamma(p) \geq \frac{(1-p)^{\gamma-2}}{p^2}\left[\gamma(1-\gamma)\,p^2\left(1-\frac{1}{p}\right) + 2\gamma\,p(1-p) + (1-p)^2\right] \tag{46}$$

$$= \frac{(1-p)^{\gamma-1}}{p^2}\left(\gamma^2\,p + \gamma\,p + 1 - p\right), \tag{47}$$

  which is always positive.

This concludes the proof of Proposition 1.

A.2 Proof of Proposition 2

The first property is directly derived from the inspection of (8).

For the second property, from (9), we have that $\mathsf{L}''_\gamma(p) > 0$ on $p \in (0,1)$. Thus, the function $p \mapsto \mathsf{L}'_\gamma(p)$ is strictly increasing on $p \in (0,1)$ and continuous. It therefore follows that $\left(\mathsf{L}'_\gamma\right)^{-1}$ is also a strictly increasing function on its domain (i.e., the range of L'_γ), that is, $t \mapsto \left(\mathsf{L}'_\gamma\right)^{-1}(t)$ is strictly increasing for $t < 0$.

Finally, the third property is a direct consequence of the first three properties.

This concludes the proof of Proposition 2.

A.3 Closed-Form Expression for the Inverse of (8)

For $\gamma = 0$ and $\gamma = 1$, we have a closed-form expression for the inverse of (8), which is given by

$$\left(\mathsf{L}'_\gamma\right)^{-1}(t) = \begin{cases} \frac{1}{\mathrm{W}_0(\mathrm{e}^{1-t})} & \gamma = 1, \\ -\frac{1}{t} & \gamma = 0. \end{cases} \tag{48}$$

A.4 Proof of Proposition 3

Property 1) follows by inspection.

To show Property 2), we start by proving that $p \mapsto \kappa(p)$ is strictly decreasing. By taking the first derivative of $\kappa(p)$ in (11), we obtain

$$\kappa'(p) = \frac{2p - 1 - \left(\log\left(\frac{1}{p}\right) - 1\right)^2}{p^2 \log^2\left(\frac{1}{p}\right)}. \tag{49}$$

The denominator is always positive and hence, we focus on studying the sign of the numerator, denoted as $f(p)$. We have that $f'(p) = \frac{2(p-\log(p)-1)}{p} > 0$, where we have used $\log\left(\frac{1}{p}\right) > 1 - p$ for all $p \in (0,1)$. Thus, $p \mapsto f(p)$ is

strictly increasing, i.e., its maximum is $f(1) = 0$. It therefore follows that $f(p) < 0$ for all $p \in (0, 1)$ and hence, $\kappa'(p) < 0$ for all $p \in (0, 1)$, that is, $p \mapsto \kappa(p)$ is strictly decreasing.

We now prove that $\kappa(p) \leq 0$ for all $p \in [\mathrm{w}, 1)$. Since $p \mapsto \kappa(p)$ is strictly decreasing, it suffices to show that $\kappa(\mathrm{w}) = 0$. Setting (11) equal to zero, we arrive at

$$2p - \log(p) - 2 = 0, \tag{50}$$

or equivalently, after some manipulation,

$$(-2p)\, \mathrm{e}^{-2p} = -2\, \mathrm{e}^{-2}, \tag{51}$$

which is satisfied for $p = \mathrm{w}$ and $p = 1$.

Now, by using the first and second derivatives of L_γ in (8) and in (9), we have that

$$\phi'(p) = \gamma\,(1-p)^{\gamma-2}\left(\log\left(\frac{1}{p}\right)(1-\gamma\, p) - 2(1-p)\right) \tag{52}$$

$$= \gamma\,(1-p)^{\gamma-2}\, p \log\left(\frac{1}{p}\right)(\kappa(p) - \gamma)\,, \tag{53}$$

which implies that the above derivative is negative whenever $\gamma > \kappa(p)$ and positive if $\gamma < \kappa(p)$.

To show Property 3), note that

$$\phi_\gamma(p) = p\,(1-p)^{\gamma-1}\left(\gamma \log\frac{1}{p} + \frac{1-p}{p}\right) \tag{54}$$

$$\leq (1-p)^{\gamma}(1+\gamma) \tag{55}$$

$$\leq 1 + \gamma, \tag{56}$$

where (55) follow by using the bound $\log x \leq x - 1$, $x > 0$.

We now show Property 4). To show the first upper bound, note that by Property 2), $\phi_\gamma(p)$ is decreasing on $p \in [\mathrm{w}, 1)$ since $\kappa(p) < 0$. Hence, $p_\gamma^+ \leq \mathrm{w}$.

The second upper bound follows by noting that p_γ^+ is such that $\kappa(p_\gamma^+) = \gamma$, which implies that

$$\frac{1}{p_\gamma^+} - \frac{2(1-p_\gamma^+)}{p_\gamma^+ \log\left(\frac{1}{p_\gamma^+}\right)} = \gamma. \tag{57}$$

By using the change of variable $p_\gamma^+ = \mathrm{e}^{-x}, x > 0$, we can rewrite the above as follows,

$$\gamma = \mathrm{e}^x - \frac{2\,(\mathrm{e}^x - 1)}{x} \leq \mathrm{e}^x - 2\,\mathrm{e}^{\frac{x}{2}}, \tag{58}$$

where the inequality follows since $\frac{\mathrm{e}^x - 1}{x} \geq \mathrm{e}^{\frac{x}{2}}$ for $x > 0$. By letting $y = \mathrm{e}^{\frac{x}{2}}$, the above inequality can be rewritten as

$$y^2 - 2y - \gamma \geq 0, \tag{59}$$

which, since $\gamma \geq 0$, implies that

$$y \geq 1 + \sqrt{1+\gamma}, \tag{60}$$

and hence, proves the upper bound on p_γ^+.

To show Property 5), we use (10) and write

$$\frac{\phi_\gamma(u)}{\phi_\gamma\left(\frac{1-u}{2}\right)} = \frac{(1-u)^\gamma\left(\gamma\frac{u}{1-u}\log\left(\frac{1}{u}\right) + 1\right)}{\left(\frac{1+u}{2}\right)^\gamma\left(\gamma\frac{1-u}{1+u}\log\left(\frac{2}{1-u}\right) + 1\right)} \tag{61}$$

$$= \left(\frac{2}{s_u}\right)^\gamma \left(\frac{1 + \gamma\frac{\log(1+t_u)}{t_u}}{1 + \gamma\frac{\log(1+s_u)}{s_u}}\right), \tag{62}$$

where we let $t_u = \frac{1-u}{u}$ and $s_u = \frac{1+u}{1-u}$. By leveraging the equality above and letting $\mathsf{T}_u = \frac{\log(1+t_u)}{t_u}$ and $\mathsf{S}_u = \frac{\log(1+s_u)}{s_u}$, we can write

$$\phi_\gamma(u) - \phi_\gamma\left(\frac{1-u}{2}\right) = \frac{\phi_\gamma\left(\frac{1-u}{2}\right)}{1+\gamma\mathsf{S}_u}\left[\left(\frac{2}{s_u}\right)^\gamma(1+\gamma\mathsf{T}_u) - (1+\gamma\mathsf{S}_u)\right]. \tag{63}$$

Our goal is to prove that the above expression is non-negative for all $u \in (0, 1/3]$. We note that from Property 1) and the fact that $\mathsf{S}_u > 0$, we have that $\frac{\phi_\gamma\left(\frac{1-u}{2}\right)}{1+\gamma\mathsf{S}_u} \geq 0$. Thus, proving that (63) is non-negative for all $u \in (0, 1/3]$ is equivalent to prove that

$$f_u(\gamma) = \left(\frac{2}{s_u}\right)^\gamma(1+\gamma\mathsf{T}_u) - (1+\gamma\mathsf{S}_u) \tag{64}$$

is non-negative for all $u \in (0, 1/3]$. We next prove that this is the case by showing that $\gamma \mapsto f_u(\gamma)$ is non-decreasing for all $u \in (0, 1/3]$, which implies that

$$f_u(\gamma) \geq f_u(0) = 0, \tag{65}$$

for all $u \in (0, 1/3]$. To show that $\gamma \mapsto f_u(\gamma)$ in (64) is non-decreasing for all $u \in (0, 1/3]$, it suffices to show that $f'_u(\gamma) \geq 0$ for all $u \in (0, 1/3]$. To prove this, we first show that $\gamma \mapsto f'_u(\gamma)$ is strictly increasing and then show that $f'_u(0) \geq 0$ for all $u \in (0, 1/3]$.

We have that

$$f'_u(\gamma) = \log\left(\frac{2}{s_u}\right)\left(\frac{2}{s_u}\right)^\gamma(1+\gamma\mathsf{T}_u) + \left(\frac{2}{s_u}\right)^\gamma\mathsf{T}_u - \mathsf{S}_u, \tag{66}$$

and

$$f''_u(\gamma) = \left(\frac{2}{s_u}\right)^\gamma\log\left(\frac{2}{s_u}\right)\left[\log\left(\frac{2}{s_u}\right)(1+\gamma\mathsf{T}_u) + 2\mathsf{T}_u\right]. \tag{67}$$

Now, since $s_u \leq 2$ and $\mathsf{T}_u > 0$ for all $u \in (0, 1/3]$, it follows that $f''_u(\gamma) > 0$ for all $u \in (0, 1/3]$. Thus, $\gamma \mapsto f'_u(\gamma)$ is strictly increasing for all $u \in (0, 1/3]$.

We now show that, for all $u \in (0, 1/3]$, it holds that $f'_u(0) \geq 0$. We have that

$$\begin{aligned} f'_u(0) &= \log\left(\frac{2}{s_u}\right) + \mathsf{T}_u - \mathsf{S}_u && (68)\\ &= \log\left(\frac{2}{s_u}\right) + \frac{\log(1+t_u)}{t_u} - \frac{\log(1+s_u)}{s_u} && (69)\\ &= \frac{1}{(1-u^2)}\left(2u(1-u)\log(2) - (1-u^2)\log(1+u) + 2(1-u)\log(1-u) - u(1+u)\log(u)\right), && (70)\end{aligned}$$

where the last equality follows by substituting the definitions of t_u and s_u. Note that the denominator of $f'_u(0)$ is always positive for all $u \in (0, 1/3]$. Therefore, to prove $f'_u(0) \geq 0$ for all $u \in (0, 1/3]$, it suffices to prove that the numerator of (70) is non-negative for all $u \in (0, 1/3]$. We denote the numerator of (70) by $g_u(0)$. In what follows, we will show that, for all $u \in (0, 1/3]$, we have that $g_u(0)$ is concave and hence, it achieves minimum value at one of its end points, which are

$$g_{0^+}(0) = g_{1/3}(0) = 0. \tag{71}$$

To prove concavity, we take the second derivative of $g_u(0)$. It is a simple exercise to obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}g_u(0) = \frac{1 - 5u^2 - 2u(1-u^2)\log\left(\frac{1+u}{4u}\right)}{u(u^2-1)}. \tag{72}$$

For all $u \in (0, 1/3]$, the denominator of the above expression is always negative and hence, to prove concavity of $g_u(0)$ it suffices to prove that the numerator of (72) is positive for all $u \in (0, 1/3]$. We denote the numerator of (72) by $h_u(0)$. We have that

$$h_u(0) = 1 - 5u^2 - 2u(1-u^2)\log\left(\frac{1+u}{4u}\right) \tag{73}$$

$$\geq 1 - 5u^2 - 2u(1-u^2)\left(\frac{1+u}{4u} - 1\right) \tag{74}$$

$$= \frac{1}{2}(1 + 3u - 9u^2 - 3u^3) \tag{75}$$

$$= \frac{1}{2}((1-9u^2) + 3u(1-u^2)) \tag{76}$$

$$> 0, \tag{77}$$

where the first inequality follows by using the fact that $\log(x) \leq x - 1$ for all $x > 0$ and the second inequality is due to the fact that, for all $u \in (0, 1/3]$, we have that $(1-9u^2) \geq 0$ and $3u(1-u^2) > 0$. Thus, for all $u \in (0, 1/3]$, $h_u(0) > 0$, which implies that $g_u(0)$ is concave with $g_u(0) \geq 0$ and hence, $f'_u(0) \geq 0$, concluding the proof that $\gamma \mapsto f_u(\gamma)$ in (64) is non-decreasing for all $u \in (0, 1/3]$. This concludes the proof of Proposition 3.

B Proofs for Section 3

B.1 Additional Properties of the Mapping $\gamma \mapsto H_\gamma(P_X, Q_X)$

Consider the mapping $\gamma \mapsto H_\gamma(P_X, Q_X)$. The first and second order derivatives are given by

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} H_\gamma(P_X, Q_X) = \mathbb{E}\left[\log(1 - Q_X(X))\mathsf{L}_\gamma(Q_X(X))\right] \leq 0, \tag{78}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\gamma^2} H_\gamma(P_X, Q_X) = \mathbb{E}\left[\left(\log(1 - Q_X(X)\right)^2 \mathsf{L}_\gamma(Q_X(X))\right] \geq 0, \tag{79}$$

where the expectations are taken with respect to $X \sim P_X$.

B.2 Proof of Proposition 5

Define a positive measure

$$\mu(x) = P_X(x) \log \frac{1}{Q_X(x)},\ x \in \mathcal{X}, \tag{80}$$

which is well-defined, in view of the fact that $P_X \ll Q_X$, and has the same support as P_X, namely $\mathcal{S}$. Moreover, μ is a finite measure since, from (2), we have that

$$\sum_x \mu(x) = H(P_X, Q_X) < \infty, \tag{81}$$

where the inequality holds from the assumption. Next, note that we can write

$$\begin{aligned} H_\gamma^{\frac{1}{\gamma}}(P_X, Q_X) &= \left(\sum_x \mu(x)(1 - Q_X(x))^\gamma\right)^{\frac{1}{\gamma}} \\ &= \|1 - Q_X(\cdot)\|_\gamma, \end{aligned} \tag{82}$$

where $\|\cdot\|_p$ is the $\ell_p(\mathcal{X}, \mu)$, $p > 0$, norm. Now, taking the limit for $\gamma \to \infty$ and using the standard result that ℓ_p norms converge to the essential supremum of the underlying function (Folland, 1999, Ch. 6, Ex. 7), we arrive at

$$\lim_{\gamma \to \infty} \|1 - Q_X(\cdot)\|_\gamma = \sup_{x \in \text{ support } \mu} (1 - Q_X(x)) \tag{83}$$

$$= \sup_{x \in \mathcal{S}} (1 - Q_X(x)), \tag{84}$$

where in the last step we have used the fact that μ is supported on $\mathcal{S}$. The proof of Proposition 5 is concluded by noting that (84) is equal to one if $\mathcal{S}$ is countably infinite.

Figure 10 provides an illustration of the convergence of $H_\gamma^{\frac{1}{\gamma}}(P_X, Q_X)$ for a fixed P_X with $|\mathcal{S}| = 3$ and two different distributions Q_X. From Figure 10, we note that this convergence is pretty fast.

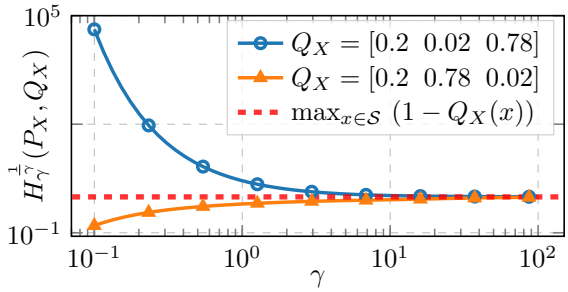

Figure 10: $P_X = \begin{bmatrix} 0.4 & 0.58 & 0.02 \end{bmatrix}$.

B.3 Proof of Proposition 6

If P_X is not absolutely continuous with respect to Q_X, the statement trivially holds by the definition in (4). Hence, suppose that $P_X \ll Q_X$. Then, on the one hand, from Proposition 4, we have that

$$H_\gamma(P_X, Q_X) \;\leq\; H(P_X, Q_X). \tag{85}$$

On the other hand, fix some $\delta > 0$ and let $\mathcal{Q}_\delta = \{x \in \mathcal{X} : Q_X(x) \leq \delta\}$. Then,

$$\begin{aligned} H_\gamma(P_X, Q_X) &\geq \sum_{x\in\mathcal{Q}_\delta} P_X(x)\,(1 - Q_X(x))^\gamma \log\left(\frac{1}{Q_X(x)}\right) \\ &\geq (1-\delta)^\gamma \sum_{x\in\mathcal{Q}_\delta} P_X(x) \log\left(\frac{1}{Q_X(x)}\right) \qquad (86) \\ &= (1-\delta)^\gamma \left(H(P_X, Q_X) - \Delta_\delta\right), \qquad (87) \end{aligned}$$

where we have defined

$$\Delta_\delta = \sum_{x\in\mathcal{Q}_\delta^c} P_X(x) \log\left(\frac{1}{Q_X(x)}\right). \tag{88}$$

Next, note that since $\mathcal{Q}_\delta^c$ is a finite set, then Δ_δ is a finite sum for all $\delta > 0$. Combining (85) and (87), we conclude that $H_\gamma(P_X, Q_X) < \infty$ if and only if $H(P_X, Q_X) < \infty$, which concludes the proof of Proposition 6.

B.4 Proof of Proposition 7

Consider a sequence Q_n that converges to a distribution Q_∞. Then, the lower semicontinuity property follows from the application of Fatou's lemma since

$$\begin{aligned} \liminf_{n\to\infty} H_\gamma(P_X, Q_n) &\geq \mathbb{E}\left[\liminf_{n\to\infty} \mathsf{L}_\gamma(Q_n(X))\right] \qquad (89) \\ &= H_\gamma(P_X, Q_\infty). \qquad (90) \end{aligned}$$

To show the convexity property, recall from Proposition 1 that $p \;\mapsto\; \mathsf{L}_\gamma(p)$ is strictly convex on $(0, 1)$ for $\gamma \geq 0$. Consequently,

$$H_\gamma(P_X, Q) = \sum_{x\in\mathcal{X}} P_X(x)\, \mathsf{L}_\gamma(Q(x)) \tag{91}$$

is a non-negative weighted sum of strictly convex functions and hence, it is convex in Q. Moreover, $H_\gamma(P_X, Q)$ is strictly convex in Q on $\mathcal{S}$ due to the fact that it is a positive weighted sum of strictly convex functions. This concludes the proof of Proposition 7.

B.5 Example: H_γ is Not Weakly Continuous

Let $\mathcal{X} = \mathbb{N}$ and consider the distribution

$$P_X(k) = 2^{-k}, \quad k = 1, 2, \ldots. \tag{92}$$

Now, for each n, define

$$Q_n(k) = \begin{cases} P_X(k) & k < n, \\ \sum_{j=n}^{\infty} P_X(j) & k = n, \\ 0 & k > n. \end{cases} \tag{93}$$

Then, for every fixed k, $Q_n(k) \to P_X(k)$ as $n \to \infty$, so Q_n converges weakly to P_X. However, the fact that $Q_n(k) = 0$ for all $k > n$ with $P_X(k) > 0$, leads to $H_\gamma(P_X, Q_n) = \infty$, whereas $H_\gamma(P_X, P_X) < \infty$. This example shows that H_γ is not weakly continuous.

B.6 Focal-Entropy vs. Cross-Entropy

We start by noting that the focal-loss in Definition 1 can be equivalently written as

$$\mathsf{L}_\gamma(P_X(x)) = \mathsf{L}_0(P_X^{(\gamma)}(x)) - h_\gamma(P_X), \tag{94}$$

where

$$P_X^{(\gamma)}(x) = \frac{P_X(x)^{(1-P_X(x))^\gamma}}{\exp\left(h_\gamma(P_X)\right)}, \tag{95}$$

and

$$h_\gamma(P_X) = \log\left(\sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma}\right). \tag{96}$$

The function $h_\gamma(P_X)$ has recently appeared in other contexts, such as in a rate distortion framework with soft-reconstruction (Dytso and Cardone, 2025). The next lemma (proof in Appendix B.7) further tightens the upper bound presented in (Dytso and Cardone, 2025) and shows that the correction term $h_\gamma(P_X)$ is uniformly bounded.

Lemma 3. *Let $P_X \in \mathcal{P}(\mathcal{X})$. Then, for $\gamma \geq 0$, the two following properties hold:*

1. *$\gamma \mapsto h_\gamma(P_X)$ is strictly increasing; and*
2. *$0 \leq h_\gamma(P_X) \leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma$.*

The next result leverages Lemma 3 and it establishes bounds on the difference between the minimum value of the cross-entropy and the minimum value of the focal-entropy.

Proposition 16. *Fix $\gamma \geq 0$. Then, it holds that*

$$0 \leq \inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} H_\gamma(P_X, Q_X) \leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma. \tag{97}$$

Proof. The lower bound follows since, from Proposition 4, we know that the mapping $\gamma \mapsto H_\gamma(P_X, Q_X)$ is non-increasing and hence,

$$\inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} H_\gamma(P_X, Q_X) \tag{98}$$

$$\geq \inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} H_0(P_X, Q_X) \tag{99}$$

$$= 0. \tag{100}$$

For the upper bound, we have that

$$\inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} H_\gamma(P_X, Q_X) \tag{101}$$

$$= \inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} \left(H\left(P_X, Q_X^{(\gamma)}\right) - h_\gamma(Q_X)\right) \tag{102}$$

$$\leq \inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} H\left(P_X, Q_X^{(\gamma)}\right) + \sup_{Q_X} h_\gamma(Q_X) \tag{103}$$

$$\leq \inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} H(P_X, Q_X) + \sup_{Q_X} h_\gamma(Q_X) \tag{104}$$

$$= \sup_{Q_X} h_\gamma(Q_X) \tag{105}$$

$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma, \tag{106}$$

where: (103) is due to the fact that $\inf_x(f(x)-g(x)) \geq \inf_x f(x) - \sup_x g(x)$, which follows since $f(x) \geq \inf_x f(x)$ and $g(x) \leq \sup_x g(x)$; (104) is due to the fact that the space over which $Q_X^{(\gamma)}$ is defined is a subset of $\mathcal{P}(\mathcal{X})$; and in (106) we have used the bound in Lemma 3. This concludes the proof of Proposition 16. □

B.7 Proof of Lemma 3

To show the first property, note that for $p \in (0,1)$, the function $\gamma \mapsto p^{(1-p)^\gamma}$ is strictly increasing and hence, $\gamma \mapsto h_\gamma(P_X)$ is strictly increasing.

To show the lower bound in the second property, note that

$$h_\gamma(P_X) \geq h_0(P_X) = \log\left(\sum_{x\in\mathcal{X}} P_X(x)\right) = 0, \tag{107}$$

where we have used the fact that $\gamma \mapsto h_\gamma(P_X)$ is strictly increasing. To prove the upper bound in the second property, we consider two cases separately:

- **Case $\gamma > 1$:** In this case, we have that

$$\sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma} \tag{108}$$

$$= \sum_{x\in\mathcal{X}} \mathrm{e}^{(1-P_X(x))^\gamma \log P_X(x)} \tag{109}$$

$$\leq \sum_{x\in\mathcal{X}} \mathrm{e}^{(1-\gamma P_X(x)) \log P_X(x)} \tag{110}$$

$$= \sum_{x\in\mathcal{X}} P_X(x)\mathrm{e}^{-\gamma P_X(x) \log P_X(x)} \tag{111}$$

$$\leq \mathrm{e}^{\frac{\gamma}{\mathrm{e}}} \sum_{x\in\mathcal{X}} P_X(x) = \mathrm{e}^{\frac{\gamma}{\mathrm{e}}}, \tag{112}$$

  where (110) follows from the Bernoulli's inequality $(1+x)^r \geq 1+rx$, $x \geq -1, r \geq 1$; and (112) follows from the bound $x\log x \geq -\frac{1}{\mathrm{e}}$. Thus, by using the above inside (96), we arrive at

$$h_\gamma(P_X) \leq \frac{\gamma}{\mathrm{e}}. \tag{113}$$

- **Case $0 \leq \gamma \leq 1$:** For this case, we let

$$f(\gamma) = \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma}, \tag{114}$$

  and we note that

$$f'(\gamma) = \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma}(1-P_X(x))^\gamma \log(P_X(x))\log(1-P_X(x)) \tag{115}$$

$$\leq \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma} \log\left(\frac{1}{P_X(x)}\right)\log\left(\frac{1}{1-P_X(x)}\right) \tag{116}$$

$$\leq \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma} \tag{117}$$

$$\leq \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))} \tag{118}$$

$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}, \tag{119}$$

where: in (116), we have used the fact that $(1-P_X(x))^\gamma \leq 1$; (117) follows from $\log(x) \leq x-1, x>0$; (118) is due to the fact that $\gamma \mapsto p^{(1-p)^\gamma}$ is strictly increasing; and in (119) we have used the bound in (112).

Now, by using the mean value theorem, there exists $c \in (0,\gamma)$ such that

$$f(\gamma) = f'(c)(\gamma - 0) + f(0) \tag{120}$$
$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma + 1, \tag{121}$$

where in the last step we have used the bound in (119) and the fact that $f(0)=1$.

Next, by using the above inside (96), we arrive at

$$h_\gamma(P_X) = \log\left(f(\gamma)\right) \tag{122}$$
$$\leq \log(\mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma + 1) \tag{123}$$
$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma, \tag{124}$$

where in the last step we have used the bound $\log(1+x) \leq x$.

Taking the maximum of both bounds concludes the proof of Lemma 3.

B.8 Proof of Corollary 1

Property 1) follows directly from the expression of $P_\gamma^\star$ in (15).

For Property 2), note that the fact that $P_X(x_1) \geq P_X(x_2)$ implies that

$$-\frac{\alpha_\gamma^\star}{P_X(x_1)} \geq -\frac{\alpha_\gamma^\star}{P_X(x_2)}. \tag{125}$$

By applying $\left(\mathsf{L}_\gamma'\right)^{-1}$ to both sides of the above inequality and recalling that $t \mapsto \left(\mathsf{L}_\gamma'\right)^{-1}(t)$ for $t \in (-\infty, 0)$ is strictly increasing (see Proposition 2), we arrive at

$$P_\gamma^\star(x_1) \geq P_\gamma^\star(x_2). \tag{126}$$

For Property 3), note that to have $P_\gamma^\star = P_X$, from (15) we would, in fact, need that

$$(\mathsf{L}_\gamma')^{-1}\left(-\tfrac{\alpha_\gamma^\star}{P_X(x)}\right) = P_X(x) \text{ for all } x \in \mathcal{S}, \tag{127}$$

or equivalently

$$\alpha_\gamma^\star = -P_X(x)\,\mathsf{L}_\gamma'\left(P_X(x)\right), \text{ for all } x \in \mathcal{S}. \tag{128}$$

Since the function $p \mapsto \mathsf{L}_\gamma'(p)$ is strictly increasing on $p \in (0,1)$ and continuous (see Proposition 1), then the above equation is satisfied if and only if $\gamma = 0$ or P_X is a uniform distribution.

This concludes the proof of Corollary 1.

B.9 Proof of Proposition 8

We start by showing (31). Toward this end, let $\phi_{\gamma,M} = \max_{t\in(0,1)} \phi_\gamma(t)$ and let

$$q_i = \left(\mathsf{L}_\gamma'\right)^{-1}\left(-\frac{\phi_{\gamma,M}}{p_i}\right), \tag{129}$$

from which we obtain

$$p_i = -\frac{\phi_{\gamma,M}}{\mathsf{L}_\gamma'(q_i)} = \frac{\phi_{\gamma,M}}{\phi_\gamma(q_i)}\,q_i \geq q_i, \tag{130}$$

for all i's. Thus, we have that

$$F(\phi_{\gamma,M}) = \sum_i (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\phi_{\gamma,M}}{p_i}\right) = \sum_i q_i \leq \sum_i p_i = 1. \tag{131}$$

Since $\alpha \mapsto F(\alpha)$ is strictly decreasing and $F(\alpha^\star_\gamma) = 1$, it follows that $\alpha^\star_\gamma \leq \phi_{\gamma,M}$. Finally, the bound $\phi_{\gamma,M} \leq 1+\gamma$ is given in Proposition 3.

We now show (32). From the definition of $\alpha^\star_\gamma$, we have that

$$\frac{1}{N} = \frac{1}{N}\sum_{x\in\mathcal{S}} (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha^\star_\gamma}{P_X(x)}\right), \tag{132}$$

and an additional note that

$$\begin{aligned}(\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha^\star_\gamma}{p_{\min}}\right) &\leq \frac{1}{N}\sum_{x\in\mathcal{S}} (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha^\star_\gamma}{P_X(x)}\right)\\ &\leq (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha^\star_\gamma}{p_{\max}}\right),\end{aligned} \tag{133}$$

where the inequalities follow from the fact that $t \mapsto (\mathsf{L}'_\gamma)^{-1}(t)$ for $t \in (-\infty, 0)$ is strictly increasing, as shown in Proposition 2. Now, using the fact that $p \mapsto \mathsf{L}'_\gamma(p)$ is strictly increasing on $p \in (0,1)$, as shown in Proposition 1, and applying it to (132) and (133) we arrive at

$$-\tfrac{\alpha^\star_\gamma}{p_{\min}} \leq \mathsf{L}'_\gamma\left(\frac{1}{N}\right) \leq -\tfrac{\alpha^\star_\gamma}{p_{\max}}, \tag{134}$$

which leads to (32).

We now prove (33). Let $p_{(i)}$ denote the ith largest entry of P_X. Under the assumption $\gamma > \kappa(p_{\min})$, we know from Proposition 3 that $\phi(\cdot)$ is strictly decreasing, i.e.,

$$\phi(p_{(1)}) < \cdots < \phi(p_{(N)}), \tag{135}$$

which, for all $i \in \{1, \ldots, N\}$, implies that

$$-\frac{\phi(p_{(1)})}{p_{(i)}} > -\frac{\phi(p_{(i)})}{p_{(i)}} = \mathsf{L}'_\gamma(p_{(i)}). \tag{136}$$

Now, by applying $\left(\mathsf{L}'_\gamma\right)^{-1}$ on both sides of the above inequality and recalling that $t \mapsto (\mathsf{L}'_\gamma)^{-1}(t)$ for $t \in (-\infty, 0)$ is strictly increasing (see Proposition 2), we get

$$p_{(i)} < (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(1)})}{p_{(i)}}\right). \tag{137}$$

Similarly, for all $i \in \{1, \ldots, N\}$, we conclude that

$$p_{(i)} > (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(N)})}{p_{(i)}}\right). \tag{138}$$

Now, evaluating $F(\cdot)$ in (18) in $\phi(p_{(1)})$, we obtain

$$F\left(\phi(p_{(1)})\right) = \sum_{i=1}^N (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(1)})}{p_{(i)}}\right) > \sum_{i=1}^N p_{(i)} = 1, \tag{139}$$

and evaluating $F(\cdot)$ in (18) in $\phi(p_{(N)})$, we obtain

$$F\left(\phi(p_{(N)})\right) = \sum_{i=1}^N (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(N)})}{p_{(i)}}\right) < \sum_{i=1}^N p_{(i)} = 1. \tag{140}$$

Now recall that $F(\alpha)$ is continuous and strictly decreasing in α. Thus, by monotonicity, we conclude the proof of (33). The proof of (34) follows by the same steps as those used for the proof of (33) with the only difference that, under the assumption $\gamma < \kappa(p_{\max})$, we know from Proposition 3 that $\phi(\cdot)$ is now strictly increasing, i.e.,

$$\phi(p_{(1)}) > \cdots > \phi(p_{(N)}), \tag{141}$$

This concludes the proof of Proposition 8.

C Proofs for Section 4

C.1 Proof of Proposition 9

Let f_0 be such that $(\mathsf{L}'_\gamma)^{-1}(f_0) = \frac{1}{|\mathcal{S}|}$. Then, using the Taylor remainder theorem, we have that

$$P^\star_\gamma(x) = (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\alpha^\star_\gamma}{P_X(x)}\right) \tag{142}$$

$$= \frac{1}{|\mathcal{S}|} - \left(\frac{\alpha^\star_\gamma}{P_X(x)} + f_0\right)\frac{\mathrm{d}}{\mathrm{d}t}(\mathsf{L}'_\gamma)^{-1}(t)|_{t=f_0} + R_{x,\infty} \tag{143}$$

$$= \frac{1}{|\mathcal{S}|} - \frac{\left(\frac{\alpha^\star_\gamma}{P_X(x)} + f_0\right)}{\mathsf{L}''_\gamma\left((\mathsf{L}'_\gamma)^{-1}(f_0)\right)} + R_{x,\infty} \tag{144}$$

$$= \frac{1}{|\mathcal{S}|} - \frac{\left(\frac{\alpha^\star_\gamma}{P_X(x)} + \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\right)}{\mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right)} + R_{x,\infty}, \tag{145}$$

where the last step follows from the fact that $f_0 = \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)$ and the remainder term is given by

$$R_{x,\infty} = \frac{\left(\frac{\alpha^\star_\gamma}{P_X(x)} + \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\right)^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\mathsf{L}'_\gamma)^{-1}(t)\Big|_{t=c} \tag{146}$$

$$= -\frac{\left(\frac{\alpha^\star_\gamma}{P_X(x)} + \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\right)^2}{2}\frac{\mathsf{L}'''_\gamma(u)}{\left(\mathsf{L}''_\gamma(u)\right)^3}\Bigg|_{u=(\mathsf{L}'_\gamma)^{-1}(c)}, \tag{147}$$

for some $c \in \left(\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right), -\frac{\alpha^\star_\gamma}{P_X(x)}\right)$. By combining the bounds on $\alpha^\star_\gamma$ in (32), it is not difficult to see that

$$R_{x,\infty} = O\left(\frac{1}{\gamma}\right). \tag{148}$$

Summing both sides of (145) over $x \in \mathcal{S}$, we arrive at

$$\alpha^\star_\gamma \sum_{x\in\mathcal{S}} \frac{1}{P_X(x)} + |\mathcal{S}|\,\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) - \mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\sum_{x\in\mathcal{S}} R_{x,\infty} = 0, \tag{149}$$

which implies that

$$\alpha^\star_\gamma = -\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\mathsf{HM}(P_X) + \frac{\mathsf{HM}(P_X)}{|\mathcal{S}|}\mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\sum_{x\in\mathcal{S}} R_{x,\infty} \tag{150}$$

$$= -\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\mathsf{HM}(P_X) + O\left(\gamma\left(1 - \frac{1}{|\mathcal{S}|}\right)^\gamma\right), \tag{151}$$

where the last step follows by using (148).

Now, inserting (151) into (145), we arrive at the conclusion that as $\gamma \to \infty$, we have that

$$P^\star_\gamma(x) = \frac{1}{|\mathcal{S}|} + \frac{\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)}{\mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right)}\left(\frac{\mathsf{HM}(P_X)}{P_X(x)} - 1\right) + O\left(\frac{1}{\gamma}\right) \tag{152}$$

$$= \frac{1}{|\mathcal{S}|} + O\left(\frac{1}{\gamma}\right), \tag{153}$$

where in the last equality we have used that $\frac{\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)}{\mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right)} = O\left(\frac{1}{\gamma}\right)$. This concludes the proof of Proposition 9.

Initial distributions used in Figure 4 and Figure 5:

$$P_X = \begin{bmatrix}0.65 & 0.35\end{bmatrix}, \text{ for } |\mathcal{S}| = 2, \tag{154}$$

$$P_X = \begin{bmatrix}0.43 & 0.32 & 0.25\end{bmatrix}, \text{ for } |\mathcal{S}| = 3, \tag{155}$$

$$P_X = \begin{bmatrix}0.35 & 0.25 & 0.25 & 0.15\end{bmatrix}, \text{ for } |\mathcal{S}| = 4. \tag{156}$$

C.2 Proof of Theorem 2

We consider two cases:

Case 1: If $p_{(i)} \in (0, p_{\gamma,a}]$ or $p_{(i)} \in [p_{\gamma,b}, 1)$, then $\alpha_\gamma^\star \geq \phi_\gamma(p_{(i)})$. Since $t \mapsto (\mathsf{L}_\gamma')^{-1}(t)$ for $t \in (-\infty, 0)$ is strictly increasing (see Proposition 2), we have that

$$p_{(i)} \geq (\mathsf{L}_\gamma')^{-1}\left(-\frac{\alpha_\gamma^\star}{p_{(i)}}\right). \tag{157}$$

From (15), the right-hand side of the above inequality is precisely $p_{(i)}^\star$ and hence, $d_i \geq 0$.

Case 2: If $p_{(i)} \in (p_{\gamma,a}, p_{\gamma,b})$, then $\alpha_\gamma^\star < \phi_\gamma(p_{(i)})$. Since $t \mapsto (\mathsf{L}_\gamma')^{-1}(t)$ for $t \in (-\infty, 0)$ is strictly increasing (see Proposition 2), this implies that

$$p_{(i)} < (\mathsf{L}_\gamma')^{-1}\left(-\frac{\alpha_\gamma^\star}{p_{(i)}}\right). \tag{158}$$

From (15), the right-hand side of the above inequality is precisely $p_{(i)}^\star$ and hence, $d_i < 0$. This concludes the proof of Theorem 2.

C.3 Proof of Proposition 10

We focus on the case $\alpha_\gamma^\star \geq 1$ (if $\alpha_\gamma^\star < 1$, then from Lemma 2, the result trivially holds). If $\alpha_\gamma^\star \geq 1$, then we need $\alpha_\gamma^\star = \phi_\gamma(p_{\gamma,b}) \geq 1$. However, we now prove that for all $p \in [1/2, 1)$, we have that $\phi_\gamma(p) < 1$, which implies that $p_{\gamma,b} < 1/2$ (to ensure that $\phi_\gamma(p_{\gamma,b}) \geq 1$) and hence, $p_{(1)} > p_{\gamma,b}$, i.e., $d_1 \geq 0$. The fact that $\phi_\gamma(p) < 1$, for all $p \in [1/2, 1)$ is because: (i) from Proposition 3, $\phi_\gamma(p)$ is unimodal and it is such that $\phi_\gamma(0^+) = 1$ and $\phi_\gamma(1) = 0$; (ii) $\phi_\gamma(1/2) < 1$; (iii) $p \mapsto \phi_\gamma(p)$ is strictly decreasing on $p \in [1/2, 1)$ (from Proposition 3). Thus, $p_{\max} \geq p_{\gamma,b}$, i.e., $d_1 \geq 0$. The fact that $d_2 < 0$ follows from Corollary 2. This concludes the proof of Proposition 10.

C.4 Proof of Proposition 11

From Proposition 3, since $p_{\gamma,a} < 1/3$, we have that

$$\phi_\gamma(p_{\gamma,a}) > \phi_\gamma\left(\frac{1 - p_{\gamma,a}}{2}\right). \tag{159}$$

Moreover, we also need

$$\alpha_\gamma^\star = \phi_\gamma(p_{\gamma,a}) = \phi_\gamma(p_{\gamma,b}). \tag{160}$$

The above two equations, together with the unimodality of ϕ_γ (see Proposition 3), imply that

$$p_{\gamma,b} < \frac{1 - p_{\gamma,a}}{2}. \tag{161}$$

We are now ready to prove that $d_1 \geq 0$. The proof is by contradiction. Assume that $d_1 < 0$. Then, from Theorem 2, we have that

$$p_{(1)} < p_{\gamma,b} < \frac{1 - p_{\gamma,a}}{2}, \tag{162}$$

which leads to

$$p_{(1)} + p_{(2)} + p_{(3)} \leq 2\,p_{(1)} + p_{(3)} \tag{163}$$

$$< 2\frac{1 - p_{\gamma,a}}{2} + p_{\gamma,a} \tag{164}$$

$$= 1, \tag{165}$$

where the second inequality follows since we need at least one sign change in the sequence $\{d_i\}, i \in \{1, 2, 3\}$ (see Corollary 2). Thus, we have found a contradiction since we need $p_{(1)} + p_{(2)} + p_{(3)} = 1$. This concludes the proof of Proposition 11.

C.5 Example: $p_{\min} < p_{\gamma,a}$

Let $\gamma = 0.2$ and $\mathcal{S} = \{x_1, x_2, x_3, x_4\}$ with

$$P_X(x) = \frac{1}{51}\begin{cases} 1 & x = x_1, \\ 20 & x = x_2, x_4, \\ 10 & x = x_3, \\ 0 & \text{otherwise.} \end{cases} \tag{166}$$

With this, from (15), we obtain

$$P_\gamma^\star = \begin{cases} 3/154 & x = x_1, \\ 195/499 & x = x_2, x_4, \\ 191/960 & x = x_3, \\ 0 & \text{otherwise.} \end{cases} \tag{167}$$

From the above, we have that

$$d_1 = d_2 = \frac{20}{51} - \frac{195}{499} = \frac{35}{25449} > 0, \tag{168}$$

$$d_3 = \frac{10}{51} - \frac{191}{960} = -\frac{47}{16320} < 0, \tag{169}$$

$$d_4 = \frac{1}{51} - \frac{3}{154} = \frac{1}{7854} > 0, \tag{170}$$

that is, $p_{\min} = \frac{1}{51} < p_{\gamma,a}$ (see also Figure 6).

C.6 Proof of Proposition 13

We start by noting that the bounds in (33), together with the fact that $\phi(p_{(i)})$ is strictly increasing in i (see Proposition 3), implies that there is a unique k_M such that

$$k_M = \max\{i : \phi(p_{(i)}) \leq \alpha_\gamma^\star\},\ k_M \in \mathbb{N}. \tag{171}$$

Now, we consider two cases:

1. **Case 1:** $i \leq k_M$. For this case, we have that $\phi(p_{(i)}) \leq \alpha_\gamma^\star$, which implies

$$-\frac{\alpha_\gamma^\star}{p_{(i)}} \leq -\frac{\phi(p_{(i)})}{p_{(i)}} = \mathsf{L}_\gamma'(p_{(i)}), \tag{172}$$

   and hence, by applying $\left(\mathsf{L}_\gamma'\right)^{-1}$ on both sides of the above inequality and recalling that $\left(\mathsf{L}_\gamma'\right)^{-1}$ is strictly increasing on its domain $(-\infty, 0)$ (see Proposition 2), we arrive at

$$p_{(i)}^\star \leq p_{(i)}, \tag{173}$$

   that is, $d_i \geq 0$.

2. **Case 2:** $i > k_M$. For this case, we have that $\phi(p_{(i)}) > \alpha_\gamma^\star$, which implies

$$-\frac{\alpha_\gamma^\star}{p_{(i)}} > -\frac{\phi(p_{(i)})}{p_{(i)}} = \mathsf{L}_\gamma'(p_{(i)}), \tag{174}$$

   and hence, by using similar steps as in Case 1, we arrive at

$$p_{(i)}^\star > p_{(i)}, \tag{175}$$

   that is, $d_i < 0$.

In summary, we have proved that $d_i \geq 0$ for all $i \leq k_M$ and $d_i < 0$ for all $i > k_M$, that is, $p_{\min} > p_{\gamma_a}$ (see also Figure 6). This concludes the proof of Proposition 13.

Remark 7. By following similar steps as above, it is not difficult to prove that, whenever $\gamma < \kappa(p_{\max})$, then $p_{\max} < p_{\gamma,b}$, that is, for these values of γ, the over-suppression regime *does* exist.

C.7 Proof of Proposition 14

We start by noting that, if $\alpha_\gamma^\star < 1$, then the result trivially holds since from Corollary 2, the sequence $\{d_i\}$ has at least one sign change. Therefore, we next assume $\alpha_\gamma^\star \geq 1$.

Using the definition of $\phi_\gamma(p)$ and the fact that $\phi_\gamma(p_{\gamma,b}) = \alpha_\gamma^\star$, we have that

$$\alpha_\gamma^\star = p_{\gamma,b}\,(1-p_{\gamma,b})^{\gamma-1}\left(\gamma \log \frac{1}{p_{\gamma,b}} + \frac{1-p_{\gamma,b}}{p_{\gamma,b}}\right) \tag{176}$$

$$\leq p_{\gamma,b}\,(1-p_{\gamma,b})^{\gamma-1}\left(\frac{\gamma}{p_{\gamma,b}} - \gamma + \frac{1-p_{\gamma,b}}{p_{\gamma,b}}\right) \tag{177}$$

$$= (1-p_{\gamma,b})^{\gamma}(\gamma+1), \tag{178}$$

where the inequality follows since $\log(x) \leq x-1$ for $x>0$. Now, since $\alpha_\gamma^\star \geq 1$, we can further bound the above as follows,

$$1 \leq (1-p_{\gamma,b})^{\gamma}(\gamma+1), \tag{179}$$

which implies

$$p_{\gamma,b} \leq 1 - \left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}}. \tag{180}$$

To show that $p_{\max} > p_{\gamma,b}$, it suffices to show that $p_{\gamma,b} < \frac{1}{|\mathcal{S}|}$ (since $p_{\max} \geq \frac{1}{|\mathcal{S}|}$). A sufficient condition for this can be obtained by setting the above upper bound to be smaller than $\frac{1}{|\mathcal{S}|}$, that is,

$$1 - \left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}} < \frac{1}{|\mathcal{S}|}, \tag{181}$$

or equivalently

$$\left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}} > \frac{|\mathcal{S}|-1}{|\mathcal{S}|}. \tag{182}$$

Since $\gamma \mapsto \left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}}$ is strictly increasing, it suffices to find $\gamma_0 > 0$ such that

$$\left(\frac{1}{\gamma_0+1}\right)^{\frac{1}{\gamma_0}} = \frac{|\mathcal{S}|-1}{|\mathcal{S}|} \tag{183}$$

to ensure that (182) is satisfied for all $\gamma > \gamma_0$. We now focus on solving (183), which is equivalent to

$$\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)^{\gamma_0} = \gamma_0 + 1, \tag{184}$$

or

$$\exp\left(\gamma_0 \log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)\right) = \gamma_0 + 1. \tag{185}$$

Letting $x = \gamma_0 + 1$ (where $x > 1$) and $n = \log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)$, we arrive at

$$\exp\left((x-1)\,n\right) = x, \tag{186}$$

or equivalently

$$\exp(-n) = x\exp(-nx), \tag{187}$$

or equivalently

$$-n\exp(-n) = -n\,x\exp(-nx). \tag{188}$$

By letting $y = -n\,x$ and $t = -n\exp(-n)$, we obtain

$$y\exp(y) = t. \tag{189}$$

Note that the above equation can be solved for y only if $t \geq -\frac{1}{\mathrm{e}}$, which in our case holds since

$$t = -n\exp(-n) \tag{190}$$
$$= -\log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)\exp\left(-\log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)\right) \tag{191}$$
$$= \frac{|\mathcal{S}|-1}{|\mathcal{S}|}\log\left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right) \tag{192}$$
$$\geq \frac{1}{2}\log\left(\frac{1}{2}\right) \tag{193}$$
$$> -\frac{1}{\mathrm{e}}, \tag{194}$$

where the inequality in (193) follows since the function in (192) is increasing in $|\mathcal{S}|$. Thus, (189) can be solved and, since $t < 0$, it has two distinct real solutions, namely

$$y = \mathrm{W}_0(t), \tag{195}$$

and

$$y = \mathrm{W}_{-1}(t). \tag{196}$$

We now show that the solution $y = \mathrm{W}_0(t)$ leads to the trivial solution $\gamma_0 = 0$. We start by noting that (195) is equivalent to

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_0\left(-n\exp(-n)\right). \tag{197}$$

Now, we know that $\mathrm{W}_0(t) \in (-1, 0)$ when $-\frac{1}{\mathrm{e}} < t < 0$ and hence, $-\frac{1}{n}\mathrm{W}_0(t) \in \left(0, \frac{1}{n}\right)$. Thus, from (197), we obtain

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_0\left(-n\exp(-n)\right) \in \left(-1, -1 + \frac{1}{n}\right). \tag{198}$$

We note that $0 \in \left(-1, -1 + \frac{1}{n}\right)$ and indeed $\gamma_0 = 0$ satisfies (184). However, we already know that $\gamma > 0$. Therefore, we only consider $y = \mathrm{W}_{-1}(t)$, which is equivalent to

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_{-1}\left(-n\exp(-n)\right), \tag{199}$$

where recall that $n = \log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)$. Note also that the right-hand side of (199) is always positive. This is because $\mathrm{W}_{-1}(t) \in (-\infty, -2\log(2))$ when $\frac{1}{2}\log\left(\frac{1}{2}\right) \leq t < 0$ and hence, $-\frac{1}{n}\mathrm{W}_{-1}(t) \in \left(\frac{2\log(2)}{n}, \infty\right)$ and $\frac{2\log(2)}{n} \geq 2$ for all $|\mathcal{S}| \geq 2$. Therefore, we conclude that $p_{\max} > p_{\gamma,b}$, whenever $\gamma > \gamma_0$. This concludes the proof of Proposition 14.

C.8 Proof of Proposition 15

We start by noting that when $p_{\min} > p_{\gamma_a}$, from Theorem 2 we have that either $p_{(i)} \in (p_{\gamma,a}, p_{\gamma,b})$ (in which case $d_i < 0$) or $p_{(i)} \in [p_{\gamma,b}, 1)$ (in which case $d_i \geq 0$) for all $i \in \{1, 2, \ldots, |\mathcal{S}|\}$. Let

$$v = \min\{i : d_i < 0\}. \tag{200}$$

Then, the sequence $\{d_i\}$ has one sign change, i.e., $d_i \geq 0$ for all $i \in \{1, 2, \ldots, v-1\}$ and $d_i < 0$ for all $i \in \{v, v+1, \ldots, |\mathcal{S}|\}$. In addition, we also know that

$$\sum_{i=1}^{|\mathcal{S}|} d_i = 0. \tag{201}$$

A vector whose entries change sign at most once (from positive to negative) and sum to zero has all partial sums non-negative, that is, for all $k \in \{1, \ldots, |\mathcal{S}|\}$, we have that

$$\sum_{i=1}^{k} d_i \geq 0, \tag{202}$$

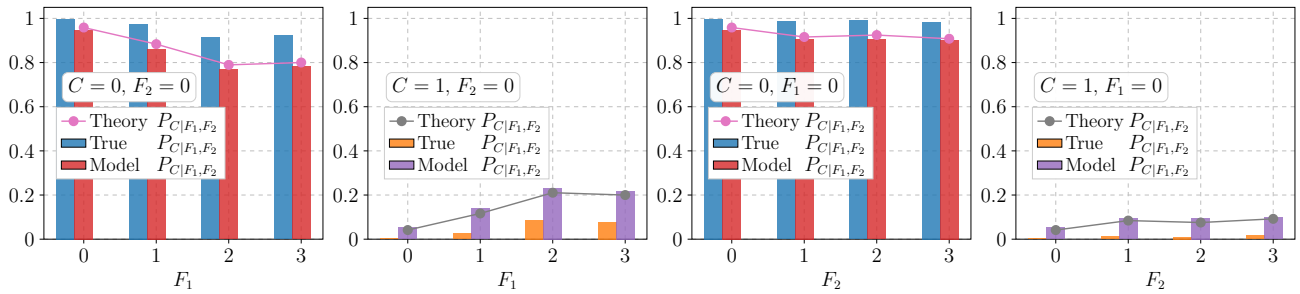

Figure 11: Synthetic dataset. For each class, we fixed one of the features at its modal value and plot: the true posterior, the posterior predicted by the trained neural network, and the theoretical focal-entropy minimizer.

or equivalently

$$\sum_{i=1}^{k} p_{(i)} \;\geq\; \sum_{i=1}^{k} p^{\star}_{(i)}, \tag{203}$$

which, together with $\sum_{i=1}^{|\mathcal{S}|} p_{(i)} = \sum_{i=1}^{|\mathcal{S}|} p^{\star}_{(i)} = 1$, is equivalent to $P_X \succ P^{\star}_{\gamma}$. This concludes the proof of Proposition 15.

D Experiments on Synthetic Data

We consider two classes $C \in \{0, 1\}$ with $P[C = 0] = 0.95$ and $P[C = 1] = 0.05$, and two features $F_1 \in \{0, 1, 2, 3\}$ and $F_2 \in \{0, 1, 2, 3\}$, Note that this is an example of an imbalanced dataset. We let

$$P[F_1|C = 0] = [0.65, 0.20, 0.10, 0.05], \tag{204}$$
$$P[F_1|C = 1] = [0.10, 0.25, 0.45, 0.20], \tag{205}$$
$$P[F_2|C = 0] = [0.50, 0.30, 0.15, 0.05], \tag{206}$$
$$P[F_2|C = 1] = [0.20, 0.50, 0.20, 0.10]. \tag{207}$$

We assume that the two features are conditionally independent given the class, i.e., $P[F_1, F_2|C] = P[F_1|C]\, P[F_2|C]$. From this, we can therefore obtain the joint probability mass function as follows,

$$P[F_1, F_2, C] = P[F_1|C]\, P[F_2|C]\, P[C], \tag{208}$$

and, by the Bayes' rule, also get $P[C|F_1, F_2]$ (see the bars labeled as "True $P_{C|F_1,F_2}$" in Figure 11 for some examples). We used this $P_{C|F_1,F_2}$ inside (15) to obtain $P^{\star}_{\gamma}$ for $\gamma = 1$, i.e., $P^{\star}_{\gamma} = \arg\min_{Q_X} H_{\gamma}(P_{C|F_1,F_2}, Q_X)$ (see the bars labeled as "Theory $P_{C|F_1,F_2}$" in Figure 11 for some examples).

We generated $10,000$ samples by repeatedly sampling from the 16 possible two-dimension feature vector space (i.e., two features with four possible values each, and two classes). These samples were used to train a feedforward neural network with a single hidden layer of 64 neurons, each using the ReLU activation function. We trained this model by using the focal loss in (3) with $\gamma = 1$ and performed the optimization with the Adam optimizer at a learning rate of 10^{-3}. Training was carried out with a batch size of 64 over 30 epochs. The trained neural network produced predicted probabilities for each of the two classes by applying the softmax function to the output layer (see the bars labeled as "Model $P_{C|F_1,F_2}$" in Figure 11 for some examples). From Figure 11, we observe that the neural network output probability is close to $P^{\star}_{\gamma}$ in (15) (i.e., with a maximum difference of 0.027). This not only supports our theoretical analysis, but it also shows that the neural network has effectively converged to the global minimum, achieving the best possible performance under the focal loss.