arXiv:2603.02533v1 [cs.IT] 3 Mar 2026

焦点熵的函数性质

Jaimin Shah
明尼苏达大学,
USA

Martina Cardone
明尼苏达大学,
USA

Alex Dytso
高通 Flarion 技术公司,
USA

摘要

焦点损失已成为类别不平衡分类问题中交叉熵的一种广泛应用的替代方案，尤其是在计算机视觉领域。尽管焦点损失在实践中取得了成功，但对其进行系统的信息论研究仍不完善。在本文中，我们采用分布视角，研究作为交叉熵对应形式的焦点熵。我们的分析确定了焦点熵具有有限性、凸性和连续性的条件，并给出了多种渐近刻画。我们证明了焦点熵极小化器的存在性与唯一性，描述了其结构，并表明它可能显著偏离数据分布。具体而言，我们严格证明，焦点损失会放大中等范围概率，抑制高概率结果，并且在类别极度不平衡时引发过度抑制区间，使得非常小的概率进一步降低。这些结果也得到了实验验证，为理解焦点损失提供了理论基础，并阐明了将其应用于不平衡学习任务时所带来的权衡。

1 引言

在基于学习的预测与决策中，损失函数塑造决策规则，并强烈影响性能。因此，选择合适的损失至关重要。一种广泛使用的选择是对数损失，

对于 $p \in (0,1]$ 定义为

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

使用分布 Q_X 进行预测时，实际结果 $x \in \mathcal{X}$ 以概率 $P_X(x)$ 发生。此时的对数损失为 $\mathsf{L}(Q_X(x)) = \log \frac{1}{Q_X(x)}$。对真实分布取平均 P_X，即可得到期望对数损失，其等于Q_X 相对于 P_X的交叉熵。形式上，$\mathcal{X}$ 上的交叉熵定义为

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

尽管对数损失应用广泛，但在类别不平衡情况下往往表现不佳；目标检测（Lin 等人，2017）、欺诈检测（He and Garcia，2009）、自然语言处理（Li 等人，2020a）以及各种医疗应用（Al Rahhal 等人，2019；Vongkulbhisal 等人，2019；Shu 等人，2019）等任务中都常见类别不平衡。为解决这一问题，Lin 等人（2017）提出了焦点损失，如今已得到广泛采用，尤其是在计算机视觉领域。它之所以成功，源于对数损失的一项简单改进：引入一个调制前置因子，降低分类正确（“容易”）样本的权重，同时提高分类错误（“困难”）样本的权重，通常能显著提升准确率。

定义1（焦点损失）。给定一个聚焦参数$\gamma \geq 0$ 和一个预测分数 $p \in (0,1]$, 焦点损失定义为

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

尽管焦点损失在实证中效果显著，但全面的信息论处理仍然缺失。交叉熵在信息论中有着清晰的理论基础——最小化交叉熵等价于最小化模型与数据之间的 Kullback–Leibler（KL）散度，从而提供了一种有理论依据的全面

29th人工智能与统计国际会议（AISTATS）2026年论文集，摩洛哥丹吉尔。PMLR：第300卷。版权所有 © 2026 作者。

优化景观。相比之下，焦点损失以尚未被充分理解的方式改变了这一景观。因此，刻画由此产生的目标函数及其极小化解的几何特性至关重要。

我们采用分布视角，并引入焦点熵，这是一种比较两个分布 P_X 和 Q_X的交叉熵的焦点损失类比形式。

定义 2（焦点熵）。 给定一个焦点参数 $\gamma \geq 0$， Q_X 相对于 P_X 在$\mathcal{X}$ 上的焦点熵定义为

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

对于 $\gamma = 0$，公式（4）中的焦点熵将退化为公式（2）中的交叉熵。交叉熵的一个关键定义性质是，在第二个参数 Q_X 上进行极小化时（在 $\mathcal{X}$上所有概率质量函数构成的空间中），其极小化解等于 P_X，即

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

其中，最后一个表达式是香农熵。正如我们在本文中所展示的，这一性质对于焦点熵不再成立，并且一般而言，优化器不等于 P_X。从这个意义上说，焦点损失可能并不表现为适当损失（Brehmer and Gneiting，2020年；Buja 等人，2005年）。然而，近期研究（Bao and Charoenphakdee，2025年）表明，焦点损失是适当评分规则的一种变体。极小化解的结构和基本性质（例如存在性），即

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

确实尚未得到充分理解。本文旨在填补这一空白，全面刻画$P_\gamma^\star$ 并研究其诸多性质。

1.1 符号说明

在集合 $\mathcal{X}$（可能是可数无限集）上所有概率质量函数构成的集合记为 $\mathcal{P}(\mathcal{X})$。对于$P_X \in \mathcal{P}(\mathcal{X})$，其支撑集为

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

我们用 $P_X \ll Q_X$ 表示 P_X 相对于 Q_X绝对连续。对于 P_X，令 $p_{\min}$ =$\min_{x\in\mathcal{S}} P_X(x)$ 和 $p_{\max}$ = $\max_{x\in\mathcal{S}} P_X(x)$。所有对数均为自然对数。朗伯函数的主分支和负分支分别记为 $\mathsf{W}_0(\cdot)$ 和 $\mathsf{W}_{-1}(\cdot)$（Corless 等人，1996年）。

1.2 文献综述

在（Charoenphakdee 等人，2021年）中，作者隐式刻画了有限字母表情形下焦点熵的极小化解。他们证明，焦点损失下的极小化类别后验分布具有保序性（即与真实后验保持相同的排序），由此得出焦点损失是 分类校准的结论。在此基础上，（Mukhoti 等人，2020年）证明，焦点损失不仅能提升困难数据集上的准确率，还能产生比交叉熵训练出的模型校准效果更好的模型。他们的分析既提供了理论依据——将焦点损失视为一种隐式最大化熵的正则化 KL 散度——也通过视觉和自然语言处理任务提供了实证证据，其中包括分布偏移下的鲁棒性。本文为他们的一项关键实证发现提供了理论依据：焦点损失的预测熵高于对数损失的预测熵。焦点损失已有多种扩展，例如修改焦点项（Li 等人，2022年，2020b）、使聚焦参数具有自适应性（Ghosh 等人，2022年），甚至证明焦点损失和对数损失是更广泛的幂级数形式的特例（Leng 等人，2022年）。相关研究方向还包括类别平衡重加权（Cui 等人，2019年）、面向长尾学习的基于间隔的方法（Cao 等人，2019年），以及平衡 Softmax 校正（Ren 等人，2020年）；这些方法从互补的角度应对类别不平衡。在校准方面，近期研究如双重焦点损失（Zhang 等人，2023年），以及倾斜交叉熵（Lin 等人，2021年）或 α-entmax 损失（Peters 等人，2019年）等基于熵的推广，进一步凸显了熵调制目标函数的核心作用。总体而言，这些进展将焦点损失置于更广泛的一类兼顾类别不平衡与校准的损失函数之中，这类函数在不同场景下权衡预测锐度、熵与鲁棒性。

1.3 论文概要与贡献

第 2 节考察焦点损失及相关函数的解析性质，重点关注其导数的逆函数等对象；该函数对于刻画焦点熵极小化器至关重要。

第 3 节阐述焦点熵的基本函数性质。命题4表明，$\gamma \mapsto H_\gamma(P_X, Q_X)$是非递增且凸的。命题 5刻画了有限支撑和可数无限支撑情况下，当 γ较大时的渐近行为。命题 6指出，焦点熵有限当且仅当交叉熵有限。命题 7证明了

$Q \mapsto H_\gamma(P_X, Q)$的弱下半连续性和凸性。最后，定理 1证明了焦点熵极小化器 $P_\gamma^\star$ 在（6）中的存在性、唯一性及结构。对于两点支撑这一特殊情形，我们还推导出关于 $P_\gamma^\star$的显式界，如命题 9所述。

第 4 节聚焦于极小化解$P_\gamma^\star$的性质，这些性质揭示了焦点损失应对类别不平衡的能力。命题 10 表明，$P_\gamma^\star$ 在 $\gamma \to \infty$时收敛于均匀分布，并给出了收敛速率。为了分析焦点熵如何将 P_X 变换为 $P_\gamma^\star$，我们研究序列 $p_{(i)} - p^\star_{(i)}$，其中$p_{(i)}$ 表示 P_X 中第 i大的元素，而 $p^\star_{(i)}$ 表示 $P_\gamma^\star$中第 i大的元素。定理 2 证明，该序列至少有一次、至多有两次符号变化。这一结果使我们能够严谨地讨论类别不平衡，包括刻画小到中等概率被放大的（即不平衡程度降低的）确切阈值。此外，我们还表明，在极端类别不平衡的情况下，也就是说某些概率非常小时，焦点损失不会放大这些概率，反而会进一步抑制它们。我们将这种情况称为过度抑制区间。这一发现对从业者尤为重要，因为它凸显了谨慎选择 γ 以避免陷入此类区间的必要性。

为更好地理解过度抑制区间，我们研究支撑集大小 γ 与数据分布 P_X 的取值之间的权衡。命题11表明，对于每个 $\gamma > 0$ 和支撑集大小 $|\mathcal{S}| = 2$，都不存在过度抑制区间。我们还推测这一结论对 $|\mathcal{S}| = 3$ 同样成立，并提供理论和数值证据支持这一猜想。命题 13 和命题 14 给出了充分条件，在这些条件下，对于任意 $|\mathcal{S}|$的取值，过度抑制区间都不存在。最后，命题 16 表明，只要过度抑制区间不存在，数据分布 P_X 就优序于极小化解 $P_\gamma^\star$。综合来看，这些结果刻画了焦点熵如何重塑数据分布。作为对这一分布视角的补充，命题17通过将焦点熵表示为相对熵，给出了信息论解释。

第 5 节在合成和真实数据（MNIST）上验证了我们的结果。最后，第 6 节总结全文。大多数证明见附录，我们实验所用的代码可在 https://tinyurl.com/3kjmmua9 获取。

2 焦点损失：解析性质

我们推导了焦点损失及相关函数的若干解析性质，这些性质将有助于

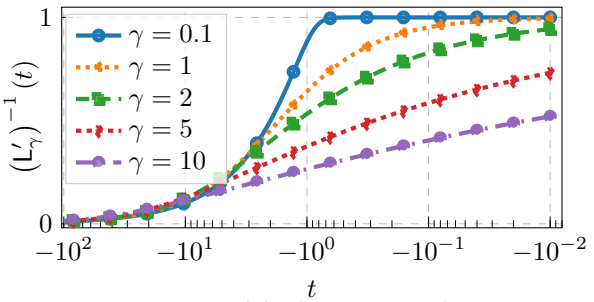


图1：式（8）中 L'_γ 的反函数。

推导本文后续章节中的结果。

在下一个命题中（证明见附录 A.1），我们考察式（3）中的焦点损失函数的单调性和凸性。

命题 1。 对于 $p \in (0,1)$ 和 $\gamma \geq 0$，有以下性质：

1. $\gamma \mapsto \mathsf{L}_\gamma(p)$ *is strictly decreasing;*
2. $p \mapsto \mathsf{L}\gamma(p)$严格递减，其导数为

$$\mathsf{L}'_\gamma(p) = -(1-p)^{\gamma-1}\left(\gamma \log\frac{1}{p} + \frac{1-p}{p}\right) < 0; \quad (8)$$

3. $p \mapsto \mathsf{L}\gamma(p)$ 是严格凸的，其二阶导数由下式给出

$$\mathsf{L}''_\gamma(p) = \frac{(1-p)^{\gamma-2}}{p^2}\left[\gamma(1-\gamma)\,p^2 \log p + 2\gamma\,p(1-p) + (1-p)^2\right] > 0. \quad (9)$$

如下文所示，L'_γ 的反函数将在求解焦点熵极小化器的一阶最优性条件中发挥重要作用。其若干重要性质总结如下（证明见附录 A.2），并在图1中加以说明。

命题 2. 对于 $p \in (0,1)$，映射 $p \mapsto \mathsf{L}'_\gamma(p)$ 具有定义良好的反函数 $(\mathsf{L}'_\gamma)^{-1} : (-\infty, 0) \to (0,1)$，并满足以下性质：

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

$(\mathsf{L}'_\gamma)^{-1}$ 的严格单调性确保映射 $t \mapsto (\mathsf{L}'_\gamma)^{-1}(t)$是单射（见图1）。对于 $\gamma = 0$ 和 $\gamma = 1$，我们可以得到式（8）的反函数的闭式表达式，如下所示：

$$(\mathsf{L}'_\gamma)^{-1}(t) = \begin{cases} \frac{1}{\mathrm{W}_0(\mathrm{e}^{1-t})} & \gamma = 1, \\ -\frac{1}{t} & \gamma = 0. \end{cases} \quad (10)$$

我们还需要最后一个函数：

$$\phi_\gamma(p) = -p\,\mathsf{L}'_\gamma(p),\ p \in (0,1). \tag{11}$$

该函数 ϕ_γ 将在第4节发挥重要作用，届时我们将研究焦点熵极小化解的性质。分析其单峰性和界，有助于解释焦点损失如何重新分配数据分布的概率质量，尤其是在类别不平衡情形下。下一个命题（证明见附录A.3）分析式（11）中 ϕ_γ的性质。

命题3. 对于所有 $\gamma \geq 0$，映射 $p \mapsto \phi_\gamma(p)$满足以下性质：

1. $\phi_\gamma(0^+) = 1$ *and* $\phi_\gamma(1) = 0$;

2. $p \mapsto \phi_\gamma(p)$ 具有单峰性。更具体地说，当 $\gamma > \kappa(p)$ 时，映射 $p \mapsto \phi_\gamma(p)$ 严格递减；当 $\gamma < \kappa(p)$ 时，该映射严格递增，其中

$$\kappa(p) = \frac{1}{p} - \frac{2(1-p)}{p\log\left(\frac{1}{p}\right)},\ p \in (0,1). \tag{12}$$

此外，

- *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. $\max\limits_{p\in(0,1)} \phi_\gamma(p) \leq 1 + \gamma$;

4. 令 $p^+_\gamma = \mathrm{argmax}_{p\in(0,1)}\,\phi_\gamma(p)$，则

$$p^+_\gamma \leq \min\left\{\mathrm{w}, \frac{1}{\left(1+\sqrt{1+\gamma}\right)^2}\right\}; \tag{13}$$

5. 对于 $u \in (0, 1/3]$，有 $\phi_\gamma(u) \geq \phi_\gamma\left(\frac{1-u}{2}\right)$.

3 焦点熵：函数性质

本节将探讨焦点熵的函数性质，并证明其凸性、有限性以及极小化解的存在性与唯一性。

3.1 焦点熵与 γ 的比较

这里我们推导出焦点熵的一些性质，以说明式（4）中的焦点熵如何随 γ 变化。第一条性质可直接从式（4）中看出（另见附录B.1，其中还给出了映射 $\gamma \mapsto H_\gamma(P_X, Q_X)$ 的一些其他性质）。

命题4。 固定某些 $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ 使得$P_X \ll Q_X$。对于 $\gamma \geq 0$，映射 $\gamma \mapsto H_\gamma(P_X, Q_X)$是非递增且凸的。因此，我们有

$$0 \leq H_\gamma(P_X, Q_X) \leq H_0(P_X, Q_X) = H(P_X, Q_X). \tag{14}$$

下一个命题（证明见附录 B.2）给出了聚焦参数 γ 趋于无穷大时焦点熵的极限。

命题 5。 取某个 $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ 使得$P_X \ll Q_X$且 $H(P_X, Q_X) < \infty$。那么，有

$$\lim_{\gamma\to\infty} H^{\frac{1}{\gamma}}_\gamma(P_X, Q_X) = \begin{cases} \max_{x\in\mathcal{S}} \bar{Q}_X(x) & |\mathcal{S}| < \infty, \\ 1 & |\mathcal{S}| = \infty, \end{cases} \tag{15}$$

where $\bar{Q}_X(x) = 1 - Q_X(x)$ *and* $\mathcal{S}$ *is defined in* (7).

我们强调，当支撑集大小 $|\mathcal{S}|$为有限值时，对 (15) 右侧关于 Q_X 进行最小化，可得到 $\mathcal{S}$上的均匀分布。关于均匀性的更精确表述将作为 $\gamma \to \infty$ 在命题 10 中给出。

3.2 基本函数性质

我们推导式 (4) 中焦点熵的一些基本函数性质。我们从下面的命题开始（证明见附录 B.3），该命题给出了焦点熵为有限值的充要条件。

命题 6。 固定某个 $\gamma \geq 0$。则，$H_\gamma(P_X, Q_X) < \infty$ 当且仅当 $H(P_X, Q_X) < \infty$。

在本文余下部分，我们仅考虑焦点熵为有限值的分布；关于熵为无限值的示例，参见（Cover 和 Thomas, 2006 年）。下一个命题（证明见附录 B.4）给出了焦点熵的其他函数性质。

命题 7. 对于任意 $\gamma \geq 0$ 和固定的 P_X，映射 $Q \mapsto H_\gamma(P_X, Q)$满足：

1. 弱下半连续[1]；并且

2. 是凸的。此外，它在定义于 $\mathcal{S}$ 上的概率空间严格 凸。见(7)。

注 1. 命题 7 表明，焦点熵是弱下半连续的。我们注意到，通常情况下，焦点熵并非弱连续。为此，令 $\mathcal{X} = \mathbb{N}$，并考虑分布

$$P_X(k) = 2^{-k}, \quad k = 1, 2, \ldots. \tag{16}$$

[1]泛函 $F\colon \mathcal{P}(\mathcal{X}) \to (-\infty, \infty]$ 称为 *weakly lower semicontinuous*，若对于任意满足 $Q_n \to Q$（弱收敛）的序列 $(Q_n)_{n\geq 1}$，都有 $F(Q) \leq \liminf_{n\to\infty} F(Q_n)$。

现在，对每个 n，定义

$$Q_n(k) = \begin{cases} P_X(k) & k < n, \\ \sum_{j=n}^{\infty} P_X(j) & k = n, \\ 0 & k > n. \end{cases} \tag{17}$$

于是，对于每个固定的 k，当 $Q_n(k) \to P_X(k)$ 且 $n \to \infty$ 时，Q_n 弱收敛于 P_X。然而，事实上，对于所有满足 $P_X(k) > 0$ 的 $k > n$，$Q_n(k) = 0$，这导致 $H_\gamma(P_X, Q_n) = \infty$，而 $H_\gamma(P_X, P_X) < \infty$。该例表明，$H_\gamma$ 不是弱连续的。

在（Mukhoti 等人，2020年）中，作者给出了焦点熵与交叉熵之间的一个不等式，从而可以得出某些学习结论。在命题 18 中，我们将给出一种新的关系，对二者下确界之差进行界定。

3.3 焦点熵的优化

式（2）中交叉熵的最小值（见式（5））源于相对熵的非负性，或等价地，源于吉布斯不等式。现在，我们将针对焦点熵证明一个类似的结果。

定理 1。固定某个 $P_X \in \mathcal{P}(\mathcal{X})$，其支撑集为 $\mathcal{S}$，如式 (7) 和 $\gamma \geq 0$ 所示。则存在唯一的 $0 < \alpha_\gamma^\star < \infty$，使得

$$P_\gamma^\star(x) = \begin{cases} (\mathsf{L}_\gamma')^{-1}\left(-\frac{\alpha_\gamma^\star}{P_X(x)}\right) & x \in \mathcal{S}, \\ 0 & x \notin \mathcal{S}, \end{cases} \tag{18}$$

是一个有效的分布。此外，对于所有 $Q_X \in \mathcal{P}(\mathcal{X})$，均有

$$H_\gamma(P_X, Q_X) \geq H_\gamma(P_X, P_\gamma^\star) + \alpha_\gamma^\star\left(1 - \sum_{x\in\mathcal{S}} Q_X(x)\right) \tag{19}$$

$$\geq H_\gamma(P_X, P_\gamma^\star), \tag{20}$$

且 $P_\gamma^\star = \arg\min_{Q_X} H_\gamma(P_X, Q_X)$ 是焦点熵的唯一极小值点。

证明。为了使 (18) 中的 $P_\gamma^\star$ 成为有效的分布，常数 $\alpha_\gamma^\star$ 必须是方程 $F(\alpha_\gamma^\star) = 1$ 的正根，其中

$$F(\alpha) = \sum_{x\in\mathcal{S}} (\mathsf{L}_\gamma')^{-1}\left(-\frac{\alpha}{P_X(x)}\right), \quad \alpha > 0. \tag{21}$$

由命题 2 可知，式 (21) 中的每一项关于 $\alpha > 0$ 都是连续且严格递减的，其取值属于 (0，1)。因此，$\alpha \mapsto F(\alpha)$ 连续且严格递减，并且 $F(0^+) = |\mathcal{S}|$ 与 $F(+\infty) = 0$（见命题 2）。于是，根据介值定理，存在唯一的正根 $\alpha_\gamma^\star$，从而得到一个有效的 $P_\gamma^\star$。

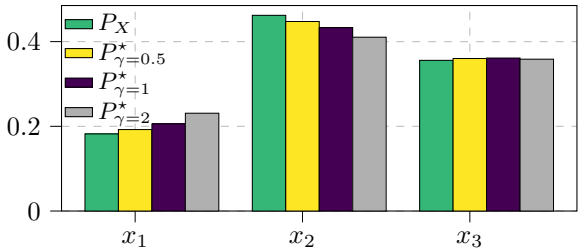


图 2：展示如何将具有 $|\mathcal{S}| = 3$ 的 P_X 转换为式 (18) 中的 $P_\gamma^\star$，其中 γ取不同值。

为证明定理的第二部分，回顾命题 1 可知，$p \mapsto \mathsf{L}_\gamma(p)$ 是严格凸的，因此可由如下切线给出下界，

$$\mathsf{L}_\gamma\left(Q(X)\right) \tag{22}$$

$$\geq \mathsf{L}_\gamma\left(P_\gamma^\star(X)\right) + \mathsf{L}_\gamma'\left(P_\gamma^\star(X)\right)\left(Q(X) - P_\gamma^\star(X)\right) \tag{23}$$

$$= \mathsf{L}_\gamma\left(P_\gamma^\star(X)\right) - \tfrac{\alpha_\gamma^\star}{P_X(X)}\left(Q(X) - P_\gamma^\star(X)\right). \tag{24}$$

因此，对于任意选择的 Q，我们得到

$$H_\gamma(P_X, Q) \tag{25}$$

$$= \mathbb{E}\left[\mathsf{L}_\gamma\left(Q(X)\right)\right] \tag{26}$$

$$\geq \mathbb{E}[\mathsf{L}_\gamma\left(P_\gamma^\star(X)\right)] - \mathbb{E}\left[\tfrac{\alpha_\gamma^\star}{P_X(X)}\left(Q(X) - P_\gamma^\star(X)\right)\right] \tag{27}$$

$$= \mathbb{E}[\mathsf{L}_\gamma\left(P_\gamma^\star(X)\right)] - \alpha_\gamma^\star\left(\sum_{x\in\mathcal{S}} Q(x) - 1\right) \tag{28}$$

$$\geq \mathbb{E}[\mathsf{L}_\gamma\left(P_\gamma^\star(X)\right)] \tag{29}$$

$$= H_\gamma(P_X, P_\gamma^\star), \tag{30}$$

其中，所有期望均相对于$X \sim P_X$取值。根据命题 7 中的凸性结果，(18) 中的 $P_\gamma^\star$是焦点熵的唯一极小化者。 □

图 2 展示了对于γ的不同取值，具有 $|\mathcal{S}| = 3$的分布 P_X 如何变换为 (18) 中的 $P_\gamma^\star$。最小的概率（位于 $x = x_1$）会增大，而最大的概率（位于 $x = x_2$）则会减小。直观地说，这种变换将概率质量从高概率结果转移到低概率结果，从而缓解不平衡。第 4 节将完整刻画哪些概率会增大或减小。

定理 1（结合命题 2）的一些直接推论将在下一个推论中给出（证明见附录 B.5）。

推论 1. 定理 *1* 中的分布 $P_\gamma^\star$ 满足以下性质：

1. $P_\gamma^\star \ll P_X$;
2. *For any* $x_1, x_2 \in \mathcal{S}$

$$P_X(x_1) \geq P_X(x_2) \implies P_\gamma^\star(x_1) \geq P_\gamma^\star(x_2); \tag{31}$$

3. $P_\gamma^\star = P_X$ 当且仅当 $\gamma = 0$ 或 P_X 为均匀分布。

注记 2. 将映射 $P_X \mapsto P_\gamma^\star$ 视为一个算子会很有帮助。该算子的反函数定义良好，且可写成如下闭式表达式：

$$P_X(x) = \frac{\beta}{\mathsf{L}'_\gamma(P_\gamma^\star(x))},\ x \in \mathcal{S}, \tag{32}$$

$$\beta = \left(\sum_{v \in \mathcal{S}} \frac{1}{\mathsf{L}'_\gamma(P_\gamma^\star(v))} \right)^{-1}. \tag{33}$$

在某些情况下，由于该逆算子具有闭式表达式，因此更易于处理。这一观点与（Charoen- phakdee 等，2021年，定理 11）相关；该文通过拉格朗日对偶，为有限字母表推导出了式 (32)。相较之下，定理 1 将这一结果推广到可数无限字母表，并基于焦点损失导数的反函数给出了更具几何直观的证明。我们的设定在实践中也更为自然：由于优化器 $P_\gamma^\star$ 通常未知，因此人们一般会使用正向映射 $P_X \mapsto P_\gamma^\star$。

式 (18) 中的极小化解取决于 $\alpha_\gamma^\star$，因此，求出 $\alpha_\gamma^\star$ 的界有助于缩小搜索空间。下一条命题（证明见附录 B.6）给出了 $\alpha_\gamma^\star$ 的界，这些界将在本文余下部分的理论推导中发挥作用。

命题8。 对于所有 $\gamma \geq 0$，均有

$$\alpha_\gamma^\star \leq \max_{t \in (0,1)} \phi_\gamma(t) \leq 1 + \gamma. \tag{34}$$

此外，假设 $|\mathcal{S}| = N < \infty$。则：

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

$$p_{\min}\ c_{N,\gamma} \leq \alpha_\gamma^\star \leq p_{\max}\ c_{N,\gamma}, \tag{35}$$

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

- 对于所有 $\gamma > \kappa(p_{\text{最小值}})$，其中 $\kappa(\cdot)$ 定义见(12)，均有

$$\phi_\gamma(p_{\max}) \leq \alpha_\gamma^\star \leq \phi_\gamma(p_{\min}). \tag{36}$$

- 对于所有 $\gamma < \kappa(p_{\text{最大值}})$，其中 $\kappa(\cdot)$ 定义见(12)，均有

$$\phi_\gamma(p_{\min}) < \alpha_\gamma^\star < \phi_\gamma(p_{\max}). \tag{37}$$

注释 3。 当 P_X 为均匀分布时，(35)、(36) 和 (37) 中的下界和上界相等。

注释4。众所周知，交叉熵的极小化解具有幂等性：若以分布 $P_0^\star$ 初始化最小化过程，则优化器与初始值一致，不会再发生变化。相比之下，焦点熵的极小化解

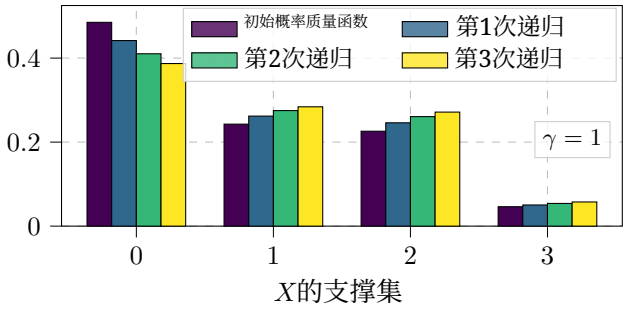


图3：初始概率质量函数：$P_X = \begin{bmatrix}0.485 & 0.243 & 0.226 & 0.046\end{bmatrix}$。

则不具备这一性质。如图3所示，式 (18) 中的优化器 $P_\gamma^\star$ 通常不具有幂等性。图3所示的结果按以下方式获得。从初始概率质量函数开始

$$P_X = \begin{bmatrix}0.485 & 0.243 & 0.226 & 0.046\end{bmatrix}, \tag{38}$$

我们通过数值计算（使用命题8中的界）确定了满足归一化条件 $F(\alpha_\gamma^\star) = 1$ 的 $\alpha_\gamma^\star$ 值，其中 $F(\cdot)$ 定义于式（21）。利用该值，我们通过式（18）（图3中的递归1）计算出相应的优化器 $P_\gamma^\star$，它与初始分布 P_X 不同。重复这一过程，将优化重新应用于新得到的分布（图3中的递归2和递归3），可以发现后续迭代仍在不断变化。因此，与交叉熵情形不同，焦点熵最小化在反复应用时会引起分布的非平凡变换。

未来一个有趣的研究方向是确定这一递归过程是否会收敛到某个极限分布。

3.4 两点支撑

我们考虑具有 $|\mathcal{S}| = 2$ 的分布，并旨在研究式（18）中优化器 $P_\gamma^\star$ 的结构。不失一般性，我们假设 $\mathcal{X} = \{0, 1\}$。下一个命题（证明见附录B.7）给出了 $P_\gamma^\star$ 的界。

命题 9. 设 $P_X \in \mathcal{P}(\{0,1\})$ 满足 $P_X(1) = p \in \left(0, \frac{1}{2}\right]$，不失一般性。则对任意 $\gamma > 0$，(18) 中在 $\mathcal{P}(\{0,1\})$ 上的唯一极小化解满足

$$\widetilde{Q}_\gamma(1) \leq P_\gamma^\star(1) \leq \widetilde{Q}_{\gamma+1}(1), \tag{39}$$

其中

$$\widetilde{Q}_\gamma(1) = q_\gamma := \frac{p^{\frac{1}{\gamma}}}{p^{\frac{1}{\gamma}} + (1-p)^{\frac{1}{\gamma}}}, \tag{40}$$

and $\widetilde{Q}_\gamma(0) = 1 - \widetilde{Q}_\gamma(1)$.

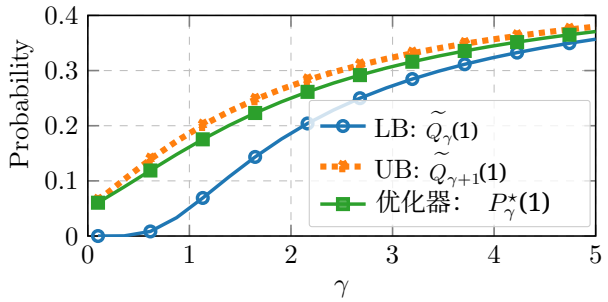


图 4：真实概率：0.05

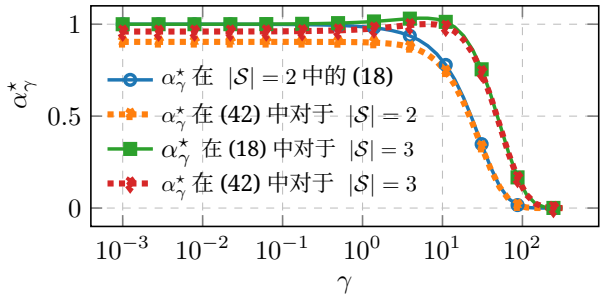


图 5： $\alpha^\star_\gamma$ 在 (18) 中与其在 (42) 中的近似的比较。

此外，对于任意 $\gamma > 0$，都有

$$\widetilde{Q}_{\gamma+1}(1) - \widetilde{Q}_\gamma(1) \leq \frac{\left|\log\left(\frac{p}{1-p}\right)\right|}{4\,\gamma^2}. \tag{41}$$

图 4 展示了命题 9 中的界在不平衡设置下的表现，即 $p = 0.05$。我们观察到，上界在 γ 的整个值域内都是紧的，而下界在 γ 取值较大时变得紧。

4 优化器的性质

在本节中，我们将探讨式 (18) 中优化器$P^\star_\gamma$的性质。这些性质至关重要，因为它们有助于揭示焦点损失的重要特征，例如它如何应对类别不平衡。

4.1 γ 较大时 $P^\star_\gamma$的渐近行为

我们的目标是了解式 (18) 中的 $P^\star_\gamma$ 如何随着 γ 增大而变化。下一个命题回答了这个问题（证明见附录 C.1）。

命题 10. 假设 $|\mathcal{S}| < \infty$。那么，当$\gamma \to \infty$时，我们有

$$\alpha^\star_\gamma = -\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\mathsf{HM}(P_X) + O\left(\gamma\left(1-\frac{1}{|\mathcal{S}|}\right)^\gamma\right), \tag{42}$$

$$P^\star_\gamma(x) = \frac{1}{|\mathcal{S}|} + O\left(\frac{1}{\gamma}\right), \tag{43}$$

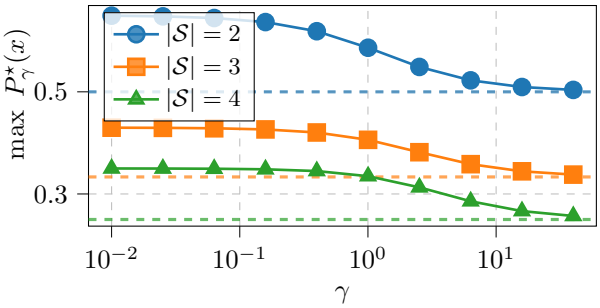


图 6： $P^\star_\gamma$ 向 $1/|\mathcal{S}|$的收敛。

其中 $\mathsf{HM}(P_X) = \frac{|\mathcal{S}|}{\sum_{x\in\mathcal{S}}\frac{1}{P_X(x)}}$ 是 P_X的调和平均数。

图 5 将式 (18) 中的 $\alpha^\star_\gamma$ 与其在式 (42) 中的近似进行比较，针对以下两种初始分布：

$$P_X = \begin{bmatrix}0.65 & 0.35\end{bmatrix}, \text{ for } |\mathcal{S}| = 2, \tag{44}$$

$$P_X = \begin{bmatrix}0.43 & 0.32 & 0.25\end{bmatrix}, \text{ for } |\mathcal{S}| = 3. \tag{45}$$

图 6 展示了对于不同的 $|\mathcal{S}|$值， $P^\star_\gamma$ 向均匀分布的收敛情况；同样， $|\mathcal{S}| = 2$和$|\mathcal{S}| = 3$ 也采用了上述相同的初始分布，并且

$$P_X = \begin{bmatrix}0.35 & 0.25 & 0.25 & 0.15\end{bmatrix}, \text{ for } |\mathcal{S}| = 4. \tag{46}$$

4.2 三分箱性质

我们希望了解何时 $P^\star_\gamma$ “支配” P_X ，反之亦然。为此，我们定义如下序列：

$$d_i = p_{(i)} - p^\star_{(i)},\ i \in \mathbb{N}, \tag{47}$$

其中，$p_{(i)}$ 是 P_X 中第 i大的元素，而 $p^\star_{(i)}$ 是式（18）中 $P^\star_\gamma$ 的第i大元素。根据推论 1， P_X和 $P^\star_\gamma$ 的排序相同；因此， d_i 定义良好，并保留了有序元素之间的对应关系。由此以及概率之和为一这一事实，可直接得到以下引理。

引理1。 如果 P_X 是均匀的，或者 $\gamma = 0$，则对于所有 $i \in \mathbb{N}$， $d_i = 0$ 。否则，序列 $\{d_i\}$ 至少发生一次变号。

在给出本节的主要结果之前（该结果可视为引理1的逆命题），我们先介绍以下预备引理。

引理2。 对于任意固定的 $\gamma > 0$， $\phi_\gamma(p) - \alpha^\star_\gamma = 0$ 至少有一个且至多有两个根。特别是，

- $p_{\gamma,a} \in [0, p^+_\gamma)$ 表示最小根，该最小根仅当 $\alpha^\star_\gamma \geq 1$ 时才存在。对于 $\alpha^\star_\gamma < 1$，我们令 $p_{\gamma,a} = 0$。
- $p_{\gamma,\,b} \in (p^+_\gamma,\ 1)$ 表示最大根，它始终存在。

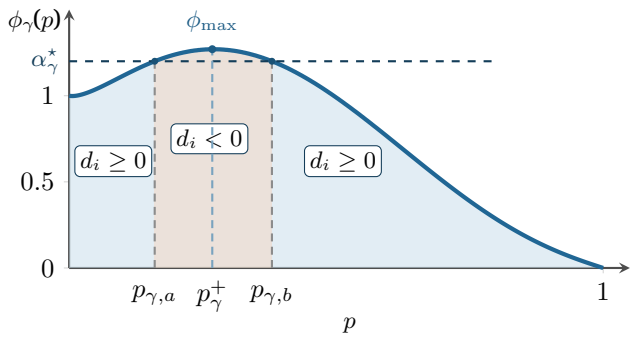


图7：定理 2 的示意图。

证明。我们回顾一下，$\phi_\gamma(p)$是单峰的（命题 3）；因此 $\phi_\gamma(p) - \alpha^\star_\gamma = 0$ 至多有两个根。由于以下原因，至少存在一个根：

$$\alpha^\star_\gamma \leq \max_{p\in(0,1)} \phi_\gamma(p), \tag{48}$$

如命题8所示。此外，由于 $\phi_\gamma(0^+)$ =1, $\phi_\gamma(1) = 0$，且 ϕ_γ 是单峰的，如果 $\alpha^\star_\gamma < 1$ 只有一个根，$p_{\gamma,\ b}$。证明至此完成。 □

下一个结果（证明见附录 C.2）刻画了 $\{d_i\}$ 中符号变化序列的行为。

定理 2. 假设 P_X 不是 均匀的，且 γ >0。那么，以下结论成立：

- *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* $d_i \geq 0$;

其中 $0 \leq p_{\gamma,a} \leq p^+_\gamma \leq p_{\gamma,b} < 1$ 在引理*2*中定义。

图7展示了定理 2 的结果。换言之，定理 2 有以下推论：

- **小到中等概率 ($p_{\gamma,a}$,$p_{\gamma,b}$)**。在此范围内，$d_i < 0$，因此，$p^\star_{(i)} > p_{(i)}$。焦点损失会放大这些概率，这是它缓解类别不平衡的主要机制，即概率质量向中等范围的条目移动。

- **大概率 [$p_{\gamma,b}$, 1)**。此处，$d_i \geq 0$，因此，$p^\star_{(i)} \leq p_{(i)}$。焦点损失会降低高概率（“容易”）样本的权重。这种效果部分源于归一化：由于中等范围概率被放大，且 $\sum_i p^\star_{(i)} = 1$，因此最大的条目必然会减小。

- **极小的概率（0，$p_{\gamma,\ a}$]**。在这个区间中，我们称之为过度抑制区间，我们还有 $d_i \geq 0$，因此，$p^\star_{(i)} \leq p_{(i)}$。

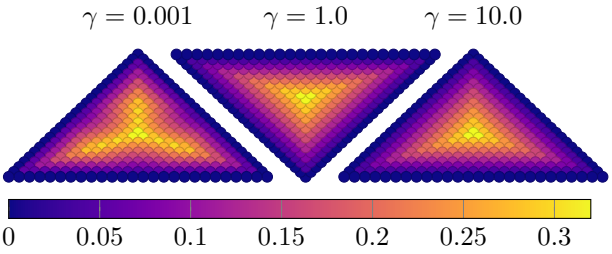


图8：$p_{\min} - p_{\gamma,\ a}$ 在三元概率单纯形上的热图。

极小的概率会受到进一步抑制，这原则上可能加剧极端尾部的不平衡。然而，如式（13）所示，当 γ 的值足够大时，区间（0，$p_{\gamma,\ a}$]为空或可以忽略。因此，尽管在实践中这种效应通常较弱，但尤其是在极端不平衡机制下，不应忽视这一区间。

定理 2 的一个直接推论见下一个推论（另见图7）。

推论 2。假设 P_X 不是均匀的，且$\gamma > 0$。那么，序列 $\{d_i\}, i \in \mathbb{N}$ 至少有一次、至多有两次符号变化。

4.3 小支撑集规模下的符号变化 es

我们证明，当支撑集为二元（证明见附录C.3），且可能为三元时，有 $p_{\min}$ >$p_{\gamma,\ a}$，即过度抑制区间永不出现。

命题11。设 $|\mathcal{S}| = 2$。那么，对于所有 $\gamma > 0$，有 $d_1 \geq 0$ 且 $d_2 < 0$。

猜想 1. 设 $|\mathcal{S}| = 3$。那么，对所有 $\gamma > 0$，均有 $p_{\min} > p_{\gamma,\ a}$，即 $d_3 < 0$（也就是说，对于 $|\mathcal{S}| = 3$，不存在过度抑制情形）。

我们现在提供支持猜想 1 的数值和理论证据。在数值方面，我们进行了大量模拟（见图8），但未能找到 $d_3 \geq 0$ 的情形。为提供理论证据，我们使用下面的结果（证明见附录C.4）。

命题12. 设 $|\mathcal{S}| = 3$。那么，对所有 $\gamma > 0$，均有 $d_1 \geq 0$。

现在，由于 $d_1 \geq 0$，可能出现两种情况：

1. $d_2 \geq 0$ （即 $p_{(2)} > p_{\gamma,\ b}$）：在这种情况下，根据推论 2，我们需要 $d_3 < 0$，即 $p_{\min} > p_{\gamma,\ a}$。因此，对于这种情况，猜想 1 成立。

2. $d_2 < 0$ （即 $p_{\gamma,\ a} < p_{(2)} < p_{\gamma,\ b}$）：对于此区间，证明 $d_3 < 0$ （即 $p_{\min} > p_{\gamma,\ a}$）是一个开放问题。换言之，这是证明猜想 1 时唯一必须处理的区间。

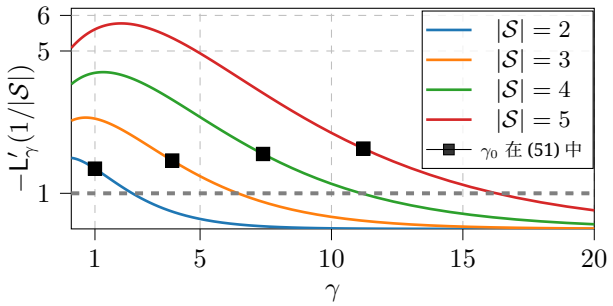


图 9：(50) 和 (51) 中条件的示意图。

注 5：尽管对于 $|\mathcal{S}| \leq 3$ 这种情况，过度抑制区间似乎对所有 $\gamma > 0$ 都不会出现，但对于 $|\mathcal{S}| \geq 4$ 则并非如此。附录 C.5给出了一个 $|\mathcal{S}| = 4$ 的例子，其中 $p_{\min} < p_{\gamma,\, a}$。

4.4 γ、支撑集基数与 P_X 之间的相互作用

我们现在着手研究 γ、支撑集基数 $\mathcal{S}$ 与 P_X 的取值之间的相互作用。特别地，我们给出了过度抑制区间不存在的充分条件。

命题 13。假设 $|\mathcal{S}| < \infty$ 且

$$-p_{\max}\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) < 1. \tag{49}$$

于是，$p_{\gamma,\, a} = 0$，也就是说，过度抑制区间不存在。

证明。结合 (35) 中对 $\alpha^\star_\gamma$ 的上界与 (49) 中的假设，可得 $\alpha^\star_\gamma < 1$。根据引理2，$\alpha^\star_\gamma < 1$ 意味着 $p_{\gamma,\, a} = 0$。至此，命题 13 的证明完成。 □

(49) 中的条件有三个自由度，即：$p_{\max}$、γ 和 $|\mathcal{S}|$。该条件还可以进一步放宽为

$$-\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) < 1. \tag{50}$$

由图 9 可知，对于每个 $|\mathcal{S}|$，都存在一个仅取决于 $|\mathcal{S}|$ 的 $\bar{\gamma}$，使得对所有 $\gamma > \bar{\gamma}$，均有 $p_{\gamma,\, a} = 0$，即过度抑制区间最终会消失。

下一个命题（证明见附录C.6）给出了 γ 最小值的一个额外下界，以确保过度抑制区间不会出现。

命题 14。设 $\kappa(\cdot)$ 由(12)给出。若 $\gamma > \kappa(p_{\min})$，则 $p_{\min} > p_{\gamma_a}$，也就是说，过度抑制区间不存在。

现在我们给出一个最终结果，说明 $p_{\max} > p_{\gamma,b}$ 的一个充分条件，即 $d_1 \geq 0$。换句话说，

下一个命题（证明见附录 C.7）确保最大概率降低。

命题 15。假设 $|\mathcal{S}| < \infty$ 和 $\gamma > \gamma_0$，其中

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_{-1}\left(-n\exp(-n)\right), \tag{51}$$

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

图 9 还报告了不同 $|\mathcal{S}|$取值下由 (51) 得到的 γ_0。命题 15 中的条件 $\gamma > \gamma_0$ 不如 (50) 严格，因为后者可确保 $d_1 \geq 0$ 和 $p_{\gamma,\, a} = 0$，而前者只能保证 $d_1 \geq 0$。

4.5 主序

已有实证观察（Mukhoti 等人，2020年）表明，焦点熵优化器 $P^\star_\gamma$ 通常具有比原始 P_X更高的熵。在此，我们针对有限支撑集大小的分布，即 $|\mathcal{S}| < \infty$，将这一观察形式化。我们从以下定义开始。

定义 3 (主序)。对于两个向量 $x, y \in \mathbb{R}^n$，若 x 优序于 y（记作 $x \succ y$），则称

$$\sum_{i=1}^{k} x_{(i)} \geq \sum_{i=1}^{k} y_{(i)}, \ \textit{for all } k = 1, \ldots, n, \tag{52}$$

其中，$x_{(i)}$ 是 x中第 i大的项。

定理 2 的一个直接推论见下一个命题（证明见附录 C.8）。

命题 16. 假设 $p_{\min} > p_{\gamma_a}$。则，P_X优序于 $P^\star_\gamma$，即 $P_X \succ P^\star_\gamma$。

关于 $p_{\min} > p_{\gamma_a}$的充分条件见命题 13 和命题 14。结合命题 16 的结果与 Schur 凸性，可得以下推论。

推论 3. 假设 $p_{\min} > p_{\gamma_a}$。则：

- Entropy increase*:*

$$H(P^\star_\gamma) \geq H(P_X), \tag{53}$$

这是因为香农熵是 *Schur* 凹的（*Kv ̊alseth, 2022*）。上述结果可推广至 *Rényi* 熵（*Kv ̊alseth, 2022*）。

- 非均匀性降低：在(53)两边同时减去对数，可 $(|\mathcal{S}|)$得

$$D_{\mathsf{KL}}(P_X \| P_U) \geq D_{\mathsf{KL}}(P^\star_\gamma \| P_U), \tag{54}$$

其中 P_U 是均匀分布。

注 6. 推论 3 将 Mukhoti 等人，2020年中的实证发现形式化为 $H(P^\star_\gamma) \geq H(P_X)$。这一点意义重大，因为熵更高的预测有助于降低过度自信（Mukhoti 等人，2020年）。

4.6 与相对熵的关系

我们的目标是将式 (4) 中的焦点熵与式 (2) 中广为研究的交叉熵联系起来。首先注意到，定义1中的焦点损失可以等价地写成

$$\mathsf{L}_\gamma(P_X(x)) = \mathsf{L}_0(P_X^{(\gamma)}(x)) - h_\gamma(P_X), \tag{55}$$

其中

$$P_X^{(\gamma)}(x) = \frac{P_X(x)^{(1-P_X(x))^\gamma}}{\exp\left(h_\gamma(P_X)\right)}, \tag{56}$$

and

$$h_\gamma(P_X) = \log\left(\sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma}\right). \tag{57}$$

函数 $h_\gamma(P_X)$ 最近也出现在其他背景中，例如软重构率失真框架（Dytso和Cardone，2025年）。下一个引理（证明见附录C.9）进一步收紧了（Dytso和Cardone，2025年）中给出的上界，并表明修正项 $h_\gamma(P_X)$ 有统一上界。

引理 3。 设 $P_X \in \mathcal{P}(\mathcal{X})$。则对于 $\gamma \geq 0$，以下两个性质成立：

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

下一个命题（证明见附录C.10）形式化地阐述了焦点熵与其他信息论量之间的联系。

命题17。 对于任意 $P_X, Q_X \in \mathcal{P}$（$\mathcal{X}$），只要满足 $P_X \ll Q_X$，就有

$$H_\gamma(P_X, Q_X) = D_{\mathsf{KL}}\left(P_X \| Q_X^{(\gamma)}\right) + H(P_X) - h_\gamma(Q_X) \tag{58}$$

$$= H_0\left(P_X, Q_X^{(\gamma)}\right) - h_\gamma(Q_X), \tag{59}$$

其中， $Q_X^{(\gamma)}$ 和 h_γ（Q_X） 分别定义于（56）和（57）。此外，还有

$$H_\gamma(P_X, Q_X) = \rho_\gamma(P_X, Q_X)\left(D_{\mathsf{KL}}(R_X \| Q_X) + H(R_X)\right), \tag{60}$$

其中

$$R_X(x) = \frac{P_X(x)(1-Q_X(x))^\gamma}{\rho_\gamma(P_X, Q_X)}, \tag{61}$$

and

$$\rho_\gamma(P_X, Q_X) = \mathbb{E}_{X\sim P_X}[(1-Q_X(X))^\gamma]. \tag{62}$$

下一个结果（证明见附录C.11）利用引理 3，确立了交叉熵最小值与焦点熵最小值之差的界。

命题 18. 固定 $\gamma \geq 0$。则有

$$0 \leq \inf_{Q_X} H(P_X, Q_X) - \inf_{Q_X} H_\gamma(P_X, Q_X) \leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma. \tag{63}$$

5 实验验证

本节将通过合成数据和真实数据验证我们的理论结果。对于合成数据的情形（附录 D），我们假设已知类别标签与特征的联合分布，从而能够直接进行理论验证。

对于真实数据，我们使用 MNIST 进行二分类： $C = 0$ 表示“数字不是 1”， $C = 1$ 表示“数字是 1”，因此存在类别不平衡（类别 0 中有 60,000 张图像中的 53,258 张）。对于每张 28×28 图像，我们提取两个用于捕捉空间不对称性的分区式特征（Impedovo 和 Pirlo，2014）： F_1，表示上半部与下半部的强度比；以及 F_2，表示左半部与右半部的强度比。每张图像由这两个特征表示，随后将其量化为四个分箱（经验分位数（Dougherty 等人，1995））。

我们使用式（3）中的焦点损失和 $\gamma = 1$ 训练了一个神经网络（架构与附录 D 中相同）。训练后的模型估计了 C 的概率，涵盖 16 种分箱组合（ F_1 分为四箱， F_2 分为四箱）——参见图 10 中的“模型 $P_{C|F_1,\ F_2}$”柱形。

由于真实后验未知，我们通过计算每种分箱组合内的类别经验比例来近似它（图 10 中的“估计$P_{C|F_1,\ F_2}$”）。将这一估计后验代入式（18），我们推导出 $\gamma = 1$ 的 $P_\gamma^\star$ （“理论 $P_{C|F_1,\ F_2}$”）。图 10 显示，网络输出与 $P_\gamma^\star$ 高度吻合（最大差异为 0.017），这支持了我们的理论，并表明在焦点损失下模型收敛至全局最小值。

6 结论

我们系统地研究了焦点熵的性质。我们证明，它保留了理想的特征（有限性和凸性），同时也表现出一些使其在根本上区别于交叉熵的行为。我们确立了其极小化解的存在性与唯一性，并描述了它如何重塑分布：放大中间范围的概率、抑制较大的概率结果，并在某些条件下过度抑制极小概率。这一发现对从业者尤为重要，因为它强调了谨慎选择 γ 的必要性，以避免进入过度抑制 区间。这些结果阐明了焦点损失带来的权衡，并为其在不平衡学习中的应用提供了理论基础。我们希望这一视角

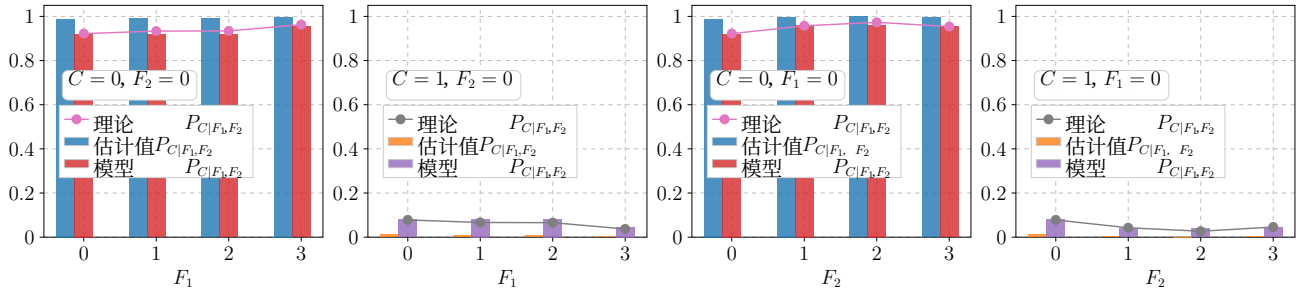


图10：MNIST数据集：对于每个类别，我们将一个特征固定在其众数值，并绘制了真实后验、训练好的神经网络预测的后验，以及理论焦点熵极小化解。“模型 $P_{C|F_1,\ F_2}$”表示对16种可能的分箱组合（即，F_1 有四个分箱，F_2也有四个分箱）中每种组合下类别 C 的预测概率，该概率由使用式（3）中的焦点损失并采用 $\gamma = 1$训练的神经网络估计（网络架构见附录D）；“估计的 $P_{C|F_1,\ F_2}$”指后验概率的经验近似（因为真实后验未知），通过计算每种分箱组合中同时具有 $C = 0$ 和 $C = 1$的样本比例得到；“理论 $P_{C|F_1,\ F_2}$”则是式（18）中的 $P^{\star}_{\gamma}$，其输入为“估计的 $P_{C|F_1,\ F_2}$”。

能为进一步的理论发展提供指引，并支持在实践中更有原则地应用焦点损失。未来一个值得探讨的方向是研究焦点损失在具有软性实值标签的分类问题中是否存在潜在益处（Jeong 等人，2023）。

A 第 2 节的证明

A.1 命题 1 的证明

为证明第一条性质，注意到对于 $p \in (0,1)$，函数 $\gamma \mapsto p^{(1-p)^\gamma}$ 严格递增，因此 $\gamma \mapsto L_\gamma(p)$ 严格递减。

通过直接计算即可得到一阶和二阶导数。显然，对于所有 $p \in (0,1)$ 和 $\gamma \geq 0$，$L'_\gamma(p) < 0$。

现在我们证明，对于所有 $p \in (0,1)$ 和 $\gamma \geq 0$，$L''_\gamma(p) > 0$，这意味着 $p \mapsto L_\gamma(p)$ 在 $(0,1)$ 上严格凸。我们分两种情况分别讨论：

- **Ca情况 $\gamma > 1$ 下 $\gamma = 0$:** </style>显然，(9) 的右侧始终为正 ;
- **情况 $0 < \gamma \leq 1$:** 利用对所有 $p > 0$ 都有 $\log(p) \geq 1 - \frac{1}{p}$ 这一事实，我们可以将 (9) 中的二阶导数下界估计如下：

$$L''_\gamma(p) \geq \frac{(1-p)^{\gamma-2}}{p^2}\left[\gamma(1-\gamma)\,p^2\left(1-\frac{1}{p}\right)+2\gamma\,p(1-p)+(1-p)^2\right] \tag{64}$$

$$= \frac{(1-p)^{\gamma-1}}{p^2}\left(\gamma^2\,p+\gamma\,p+1-p\right), \tag{65}$$

其始终为正。

命题 1 的证明至此完成。

A.2 命题 2 的证明

第一条性质可直接由检查式 (8) 得出。

对于第二条性质，由 (9) 可得，$L''_\gamma(p) > 0$ 在 $p \in (0,1)$ 上。因此，函数 $p \mapsto L'_\gamma(p)$ 在 $p \in (0,1)$ 上严格递增且连续。由此可知，$\left(L'_\gamma\right)^{-1}$ 在其定义域（即 L'_γ 的值域）上也是严格递增函数，也就是说，对于 $t < 0$，$t \mapsto \left(L'_\gamma\right)^{-1}(t)$ 严格递增。

最后，第三条性质是前三条性质的直接推论。

命题 2 的证明至此完成。

A.3 命题3 的证明

性质 1) 显而易见。

为证明性质 2)，我们首先证明 $p \mapsto \kappa(p)$ 严格递减。对式 (12) 中的$\kappa(p)$ 求一阶导数，得到

$$\kappa'(p) = \frac{2p-1-\left(\log\left(\frac{1}{p}\right)-1\right)^2}{p^2\log^2\left(\frac{1}{p}\right)}. \tag{66}$$

分母始终为正，因此我们重点考察分子的符号，记为 $f(p)$。我们有 $f'(p) = {}^{2}{}_{(p-\text{对数}(p)-1)}{}_{p} > 0$，其中用到了对于所有 $p \in (0,1)$，~~log~~ $\left(\frac{1}{p}\right) > 1-p$。因此，$p \mapsto f(p)$ 严格递增，即其最大值为 $f(1) = 0$。由此可知，对于所有 $p \in (0,1)$，$f(p) < 0$，进而对于所有 $p \in (0,1)$，$\kappa'(p) < 0$，也就是说，$p \mapsto \kappa(p)$ 严格递减。

现在我们证明，对于所有 $p \in [w,\ 1)$，$\kappa(p) \leq 0$。由于 $p \mapsto \kappa(p)$ 严格递减，只需证明$\kappa(w) = 0$。令 (12) 等于零，可得

$$2p - \log(p) - 2 = 0, \tag{67}$$

或者经过一些变形后，等价地可得，

$$(-2p)\,\mathrm{e}^{-2p} = -2\,\mathrm{e}^{-2}, \tag{68}$$

这对 $p = \mathsf{w}$ 和 $p = 1$ 均成立。

现在，利用 (8) 和 (9) 中 L_γ 的一阶和二阶导数，可得

$$\phi'(p) = \gamma\,(1-p)^{\gamma-2}\left(\log\left(\frac{1}{p}\right)(1-\gamma\,p) - 2(1-p)\right) \tag{69}$$

$$= \gamma\,(1-p)^{\gamma-2}\,p\log\left(\frac{1}{p}\right)(\kappa(p)-\gamma)\,, \tag{70}$$

这意味着，当 $\gamma > \kappa(p)$ 时，上述导数为负；而当 $\gamma < \kappa(p)$ 时，该导数为正。为证明性质 3)，
注意到

$$\phi_\gamma(p) = p\,(1-p)^{\gamma-1}\left(\gamma\log\frac{1}{p} + \frac{1-p}{p}\right) \tag{71}$$

$$\leq (1-p)^\gamma(1+\gamma) \tag{72}$$

$$\leq 1+\gamma, \tag{73}$$

其中，(72) 可由利用界 $\log x \leq x-1,\, x>0$ 得出。

现在证明性质 4)。为证明第一个上界，注意到根据性质 2)， $\phi_\gamma(p)$ 在$p \in [\mathsf{w},1)$ 上递减，因为 $\kappa(p) < 0$。因此，$p_\gamma^+ \leq \mathsf{w}$。

第二　　个上界可由以下观察得出： p_γ^+满足 $\kappa(p_\gamma^+) = \gamma$，这意味着

$$\frac{1}{p_\gamma^+} - \frac{2(1-p_\gamma^+)}{p_\gamma^+\log\left(\frac{1}{p_\gamma^+}\right)} = \gamma. \tag{74}$$

通过变量替换 $p_\gamma^+ = \mathsf{e}^{-x}, x>0$，我们可以将上式改写如下：

$$\gamma = \mathrm{e}^x - \frac{2\,(\mathrm{e}^x-1)}{x} \leq \mathrm{e}^x - 2\,\mathrm{e}^{\frac{x}{2}}, \tag{75}$$

其中，该不等式成立是因为对于 $x>0$，$\mathrm{e}^x = 1 \geq \mathsf{e}^{\frac{x}{2}}$。令 $y = \mathsf{e}^{\frac{x}{2}}$，则上述不等式可改写为

$$y^2 - 2y - \gamma \geq 0, \tag{76}$$

由于 $\gamma \geq 0$，这意味着

$$y \geq 1 + \sqrt{1+\gamma}, \tag{77}$$

因此，得出了关于 p_γ^+ 的上界。

为证明性质 5)，我们使用 (11) 并写出

$$\frac{\phi_\gamma(u)}{\phi_\gamma\left(\frac{1-u}{2}\right)} = \frac{(1-u)^\gamma\left(\gamma\frac{u}{1-u}\log\left(\frac{1}{u}\right)+1\right)}{\left(\frac{1+u}{2}\right)^\gamma\left(\gamma\frac{1-u}{1+u}\log\left(\frac{2}{1-u}\right)+1\right)} \tag{78}$$

$$= \left(\frac{2}{s_u}\right)^\gamma\left(\frac{1+\gamma\frac{\log(1+t_u)}{t_u}}{1+\gamma\frac{\log(1+s_u)}{s_u}}\right), \tag{79}$$

其中我们令 $t_u = \frac{1-u}{u}$ 和 $s_u = \frac{1+u}{1-u}$。利用上述等式，并令 $\mathsf{T}_u = \text{对数}(1+t_u)\ t_u$ 和 $\mathsf{S}_u = \text{对数}(1+s_u)\ s_u$，我们可以写出

$$\phi_\gamma(u) - \phi_\gamma\left(\frac{1-u}{2}\right) = \frac{\phi_\gamma\left(\frac{1-u}{2}\right)}{1+\gamma\mathsf{S}_u}\left[\left(\frac{2}{s_u}\right)^\gamma(1+\gamma\mathsf{T}_u) - (1+\gamma\mathsf{S}_u)\right]. \tag{80}$$

我们的目标是证明上述表达式对所有 $u \in (0, 1/3]$ 均为非负。我们注意到，由性质 1) 以及 $\mathsf{S}_u > 0$ 这一事实，可得 $\frac{\phi_\gamma(\frac{1-u}{2})}{1+\gamma \mathsf{S}_u} \geq 0$。因此，证明 (80) 对所有 $u \in (0, 1/3]$ 均为非负，等价于证明

$$f_u(\gamma) = \left(\frac{2}{s_u}\right)^\gamma (1 + \gamma \mathsf{T}_u) - (1 + \gamma \mathsf{S}_u) \tag{81}$$

对所有 $u \in (0,\ 1/3]$ 均非负。接下来，我们通过证明对所有 $u \in (0,\ 1/3]$，$\gamma \mapsto f_u(\gamma)$ 非递减，来证明这一点，这意味着

$$f_u(\gamma) \geq f_u(0) = 0, \tag{82}$$

对所有 $u \in (0,\ 1/3]$ 均成立。为了证明式 (81) 中的 $\gamma \mapsto f_u(\gamma)$ 对所有 $u \in (0,\ 1/3]$ 均非递减，只需证明对所有 $u \in (0,\ 1/3]$，$f_u'(\gamma) \geq 0$。为此，我们先证明 $\gamma \mapsto f_u'(\gamma)$ 严格递增，然后证明对所有 $u \in (0,\ 1/3]$，$f_u'(0) \geq 0$。

我们有

$$f_u'(\gamma) = \log\left(\frac{2}{s_u}\right)\left(\frac{2}{s_u}\right)^\gamma (1 + \gamma \mathsf{T}_u) + \left(\frac{2}{s_u}\right)^\gamma \mathsf{T}_u - \mathsf{S}_u, \tag{83}$$

and

$$f_u''(\gamma) = \left(\frac{2}{s_u}\right)^\gamma \log\left(\frac{2}{s_u}\right)\left[\log\left(\frac{2}{s_u}\right)(1 + \gamma \mathsf{T}_u) + 2\mathsf{T}_u\right]. \tag{84}$$

现在，由于对所有 $u \in (0,\ 1/3]$，$s_u \leq 2$ 且 $\mathsf{T}_u > 0$，可知对所有 $u \in (0,\ 1/3]$，$f_u''(\gamma) > 0$。因此，对所有 $u \in (0,\ 1/3]$，$\gamma \mapsto f_u'(\gamma)$ 严格递增。

现在我们证明，对于所有 $u \in (0, 1/3]$，都有 $f_u'(0) \geq 0$。我们有

$$\begin{aligned} f_u'(0) &= \log\left(\frac{2}{s_u}\right) + \mathsf{T}_u - \mathsf{S}_u && (85)\\ &= \log\left(\frac{2}{s_u}\right) + \frac{\log(1+t_u)}{t_u} - \frac{\log(1+s_u)}{s_u} && (86)\\ &= \frac{1}{(1-u^2)}\left(2u(1-u)\log(2) - (1-u^2)\log(1+u) + 2(1-u)\log(1-u) - u(1+u)\log(u)\right), && (87) \end{aligned}$$

其中，最后一个等式是将 t_u 和 s_u 的定义代入后得到的。注意，对于所有 $u \in (0, 1/3]$，$f_u'(0)$ 的分母始终为正。因此，要证明对于所有 $u \in (0, 1/3]$，$f_u'(0) \geq 0$，只需证明对于所有 $u \in (0, 1/3]$，(87) 的分子非负。我们将 (87) 的分子记为 $g_u(0)$。接下来，我们将证明，对于所有 $u \in (0, 1/3]$，$g_u(0)$ 是凹函数，因此它在端点之一处取得最小值，这两个端点为

$$g_{0^+}(0) = g_{1/3}(0) = 0. \tag{88}$$

为证明凹性，我们对 $g_u(0)$ 求二阶导数。通过简单计算可得

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2} g_u(0) = \frac{1 - 5u^2 - 2u(1-u^2)\log\left(\frac{1+u}{4u}\right)}{u(u^2-1)}. \tag{89}$$

对于所有 $u \in (0,\ 1/3]$，上述表达式的分母始终为负，因此，要证明 $g_u(0)$ 的凹性，只需证明对于所有 $u \in (0,\ 1/3]$，(89) 的分子为正。我们将 (89) 的分子记为 $h_u(0)$。可得

$$\begin{aligned} h_u(0) &= 1 - 5u^2 - 2u(1-u^2)\log\left(\frac{1+u}{4u}\right) && (90)\\ &\geq 1 - 5u^2 - 2u(1-u^2)\left(\frac{1+u}{4u} - 1\right) && (91)\\ &= \frac{1}{2}(1 + 3u - 9u^2 - 3u^3) && (92)\\ &= \frac{1}{2}((1 - 9u^2) + 3u(1-u^2)) && (93) \end{aligned}$$

$$> 0, \tag{94}$$

其中，第一个不等式是利用对数$(x) \leq x-1$对所有 $x>0$成立这一事实得到的；第二个不等式则是因为对于所有 $u \in (0,\ 1/3]$，有$(1-9u^2) \geq 0$且 $3u(1-u^2) > 0$。因此，对于所有 $u \in (0,\ 1/3]$，$h_u(0) > 0$，这意味着 $g_u(0)$是凹函数，且 $g_u(0) \geq 0$，因此 $f_u'(0) \geq 0$。由此证明了，对于所有 $u \in (0,\ 1/3]$，(81) 中的$\gamma \mapsto f_u(\gamma)$是非递减的。命题3的证明至此结束。

B 第 3 节的证明

B.1 映射 $\gamma \mapsto H_\gamma(P_X, Q_X)$ 的其他性质

考虑映射 $\gamma \mapsto H_\gamma(P_X, Q_X)$。其一阶和二阶导数为

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} H_\gamma(P_X, Q_X) = \mathbb{E}\left[\log(1-Q_X(X))\mathsf{L}_\gamma(Q_X(X))\right] \leq 0, \tag{95}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\gamma^2} H_\gamma(P_X, Q_X) = \mathbb{E}\left[\left(\log(1-Q_X(X)\right)^2 \mathsf{L}_\gamma(Q_X(X))\right] \geq 0, \tag{96}$$

其中，期望是相对于 $X \sim P_X$ 取的。

B.2 命题 5 的证明

定义一个正测度

$$\mu(x) = P_X(x) \log \frac{1}{Q_X(x)},\ x \in \mathcal{X}, \tag{97}$$

鉴于 $P_X \ll Q_X$，该测度是良定义的，并且与 P_X具有相同的支撑集，即 $\mathcal{S}$。此外，μ是一个有限测度，因为由(2)可得

$$\sum_x \mu(x) = H(P_X, Q_X) < \infty, \tag{98}$$

其中该不等式由假设成立。接下来，注意到我们可以写成

$$\begin{aligned} H_\gamma^{\frac{1}{\gamma}}(P_X, Q_X) &= \left(\sum_x \mu(x)(1-Q_X(x))^\gamma\right)^{\frac{1}{\gamma}} \\ &= \|1-Q_X(\cdot)\|_\gamma, \end{aligned} \tag{99}$$

其中 $\|\cdot\|_p$是 $\ell_p(\mathcal{X}, \mu)$ $p>0,$范数。现在，对 $\gamma \to \infty$取极限，并利用 ℓ_p范数收敛于相应函数的本质上确界这一标准结果（Folland，1999，第6章，习题7），我们得到

$$\lim_{\gamma\to\infty} \|1-Q_X(\cdot)\|_\gamma = \sup_{x \in \text{ support } \mu} (1-Q_X(x)) \tag{100}$$

$$= \sup_{x\in\mathcal{S}} (1-Q_X(x)), \tag{101}$$

其中在最后一步我们利用了 μ 支撑在 $\mathcal{S}$上这一事实。注意到若 $\mathcal{S}$ 为可数无穷，则(101)等于1，由此完成命题 5的证明。

[1]图 11给出了固定 P_X（满足 $|\mathcal{S}|=3$）以及两个不同分布 Q_X时 $H\gamma(P_X, Q_X)$收敛情况的示意图。从图 11可以看出，这种收敛相当快。

B.3 命题 6 的证明

如果 P_X 不关于 Q_X绝对连续，那么根据式（4）中的定义，该结论显然成立。因此，假设 $P_X \ll Q_X$。那么，一方面，由命题4可知

$$H_\gamma(P_X, Q_X) \ \leq\ H(P_X, Q_X). \tag{102}$$

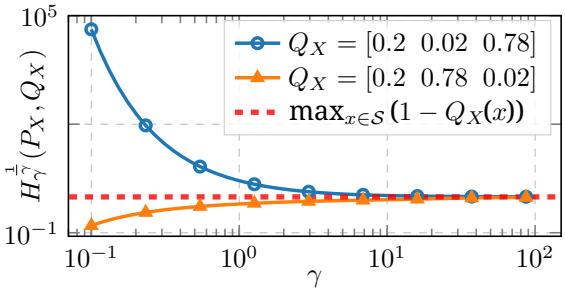


图 11: $P_X = [0.4 \quad 0.58 \quad 0.02]$。

另一方面，固定某个 $\delta > 0$，并令 $\mathcal{Q}_\delta = \{x \in \mathcal{X} \colon Q_X(x) \leq \delta\}$。那么，

$$\begin{aligned} H_\gamma(P_X, Q_X) &\geq \sum_{x\in\mathcal{Q}_\delta} P_X(x)\,(1 - Q_X(x))^\gamma \log\left(\frac{1}{Q_X(x)}\right) \\ &\geq (1-\delta)^\gamma \sum_{x\in\mathcal{Q}_\delta} P_X(x) \log\left(\frac{1}{Q_X(x)}\right) \qquad (103) \\ &= (1-\delta)^\gamma \left(H(P_X, Q_X) - \Delta_\delta\right), \qquad (104) \end{aligned}$$

其中我们定义了

$$\Delta_\delta = \sum_{x\in\mathcal{Q}_\delta^c} P_X(x) \log\left(\frac{1}{Q_X(x)}\right). \qquad (105)$$

接下来，注意到由于 $\mathcal{Q}_\delta^c$ 是有限集，因此对于所有 $\delta > 0$，Δ_δ 都是有限和。结合（102）和（104），我们得出 $H_\gamma(P_X, Q_X) < \infty$ 当且仅当 $H(P_X, Q_X) < \infty$，由此完成命题 6 的证明。

B.4 命题 7 的证明

考虑一个收敛到分布 Q_∞ 的序列 Q_n。由法图引理可得下半连续性性质，因为

$$\begin{aligned} \liminf_{n\to\infty} H_\gamma(P_X, Q_n) &\geq \mathbb{E}\left[\liminf_{n\to\infty} \mathsf{L}_\gamma(Q_n(X))\right] \qquad (106) \\ &= H_\gamma(P_X, Q_\infty). \qquad (107) \end{aligned}$$

为证明凸性，回顾命题 1 可知，对于 $\gamma \geq 0$，$p \mapsto \mathsf{L}\gamma(p)$ 在 (0,1) 上严格凸。因此，

$$H_\gamma(P_X, Q) = \sum_{x\in\mathcal{X}} P_X(x)\,\mathsf{L}_\gamma(Q(x)) \qquad (108)$$

它是严格凸函数的非负加权和，因此关于 Q 是凸的。此外，由于 $H_\gamma(P_X, Q)$ 是严格凸函数的正加权和，因此它关于 Q 在 $\mathcal{S}$ 上严格凸。至此，命题 7 的证明完成。

B.5 推论 1 的证明

性质 1) 可直接由式 (18) 中 $P_\gamma^\star$ 的表达式得出。

对于性质 2），请注意 $P_X(x_1) \geq P_X(x_2)$这一事实意味着

$$-\frac{\alpha_\gamma^\star}{P_X(x_1)} \geq -\frac{\alpha_\gamma^\star}{P_X(x_2)}. \qquad (109)$$

对上述不等式两边应用 $\left(\mathsf{L}'_\gamma\right)^{-1}$，并回想起当 $t \in (-\infty, 0)$ 时，$t \mapsto \left(\mathsf{L}'_\gamma\right)^{-1}(t)$ 严格递增（见命题 2），可得

$$P_\gamma^\star(x_1) \geq P_\gamma^\star(x_2). \qquad (110)$$

对于性质 3），注意到要使 $P_\gamma^\star = P_X$ 成立，根据（18），实际上需要满足

$$(\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha_\gamma^\star}{P_X(x)}\right) = P_X(x) \text{ for all } x \in \mathcal{S}, \tag{111}$$

或者等价地

$$\alpha_\gamma^\star = -P_X(x)\,\mathsf{L}'_\gamma\left(P_X(x)\right), \text{ for all } x \in \mathcal{S}. \tag{112}$$

由于函数 $p \mapsto \mathsf{L}'_\gamma(p)$ 在 $p \in (0,1)$ 上严格递增且连续（见命题 1），因此上述等式成立，当且仅当 $\gamma = 0$ 或 P_X 是均匀分布。

推论 1 的证明至此完成。

B.6 命题8的证明

我们首先证明 (34)。为此，令 $\phi_{\gamma,M} = \max_{t\in(0,1)} \phi_\gamma(t)$，并令

$$q_i = \left(\mathsf{L}'_\gamma\right)^{-1}\left(-\frac{\phi_{\gamma,M}}{p_i}\right), \tag{113}$$

由此我们得到

$$p_i = -\frac{\phi_{\gamma,M}}{\mathsf{L}'_\gamma(q_i)} = \frac{\phi_{\gamma,M}}{\phi_\gamma(q_i)}\,q_i \geq q_i, \tag{114}$$

对所有的 i 都成立。因此，我们有

$$F(\phi_{\gamma,M}) = \sum_i (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\phi_{\gamma,M}}{p_i}\right) = \sum_i q_i \leq \sum_i p_i = 1. \tag{115}$$

由于 $\alpha \mapsto F(\alpha)$ 严格递减，且 $F(\alpha_\gamma^\star) = 1$，因此可得 $\alpha_\gamma^\star \leq \phi_{\gamma,\,M}$。最后，界 $\phi_{\gamma,\,M} \leq 1+\gamma$由命题3给出。

现在证明 (35)。根据 $\alpha_\gamma^\star$的定义，我们有

$$\frac{1}{N} = \frac{1}{N}\sum_{x\in\mathcal{S}} (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha_\gamma^\star}{P_X(x)}\right), \tag{116}$$

并补充说明

$$\begin{aligned}(\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha_\gamma^\star}{p_{\min}}\right) &\leq \frac{1}{N}\sum_{x\in\mathcal{S}} (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha_\gamma^\star}{P_X(x)}\right) \\ &\leq (\mathsf{L}'_\gamma)^{-1}\left(-\tfrac{\alpha_\gamma^\star}{p_{\max}}\right),\end{aligned} \tag{117}$$

其中，不等式由以下事实得出：对于 $t \in (-\infty,\ 0)$，$t \mapsto (\mathsf{L}'_\gamma)^{-1}(t)$ 严格递增，正如命题 2 所示。现在，利用命题 1 所示的事实：$p \mapsto \mathsf{L}'_\gamma(p)$ 在 $p \in (0,\ 1)$ 上严格递增，并将其应用于 (116) 和 (117)，可得

$$-\tfrac{\alpha_\gamma^\star}{p_{\min}} \leq \mathsf{L}'_\gamma\left(\frac{1}{N}\right) \leq -\tfrac{\alpha_\gamma^\star}{p_{\max}}, \tag{118}$$

由此可得(35)。

我们现在证明(36)。令$p_{(i)}$ 表示 P_X的第 i大项。在假设 $\gamma > \kappa(p_{\min})$下，由命题3可知，$\phi(\cdot)$是严格递减的，即

$$\phi(p_{(1)}) < \cdots < \phi(p_{(N)}), \tag{119}$$

这对所有 $i \in \{1,\ldots,N\}$ 都意味着

$$-\frac{\phi(p_{(1)})}{p_{(i)}} > -\frac{\phi(p_{(i)})}{p_{(i)}} = \mathsf{L}'_\gamma(p_{(i)}). \tag{120}$$

现在，对上述不等式两边应用 $\left(\mathsf{L}'_\gamma\right)^{-1}$，并回想起当 $t \in (-\infty, 0)$ 时，$t \mapsto (\mathsf{L}'_\gamma)^{-1}(t)$ 严格递增（见命题 2），我们得到

$$p_{(i)} < (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(1)})}{p_{(i)}}\right). \tag{121}$$

类似地，对所有 $i \in \{1, \ldots, N\}$，我们得出

$$p_{(i)} > (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(N)})}{p_{(i)}}\right). \tag{122}$$

现在，在 $\phi(p_{(1)})$处计算(21)中的 $F(\cdot)$，我们得到

$$F\left(\phi(p_{(1)})\right) = \sum_{i=1}^{N} (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(1)})}{p_{(i)}}\right) > \sum_{i=1}^{N} p_{(i)} = 1, \tag{123}$$

以及在 $\phi(p_{(N)})$处计算(21)中的 $F(\cdot)$，我们得到

$$F\left(\phi(p_{(N)})\right) = \sum_{i=1}^{N} (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\phi(p_{(N)})}{p_{(i)}}\right) < \sum_{i=1}^{N} p_{(i)} = 1. \tag{124}$$

现在回顾，$F(\alpha)$关于 α连续且严格递减。因此，由单调性，我们完成了(36)的证明。(37)的证明遵循与(36)的证明相同的步骤，唯一的区别是，在假设 $\gamma < \kappa(p_{\max})$下，由命题3可知 $\phi(\cdot)$现在是严格递增的，即

$$\phi(p_{(1)}) > \cdots > \phi(p_{(N)}), \tag{125}$$

至此完成命题8的证明。

B.7 命题 9 的证明

根据命题 7，只要分布 P_X 非退化，映射 $Q_X \mapsto H_\gamma(P_X, Q_X)$ 在 $\{0, 1\}$ 上就是严格凸的。因此，令 $q = Q_X(1)$（或等价地，令 $Q_X(0) = 1 - q$），则导数

$$D(q) = \frac{\mathrm{d}}{\mathrm{d}q} H_\gamma(P_X, \{q, 1-q\}) \tag{126}$$

$$= p\,(1-q)^{\gamma-1}\left(\gamma \log(q) - \frac{1-q}{q}\right) + (1-p)\,q^{\gamma-1}\left(-\gamma \log(1-q) + \frac{q}{1-q}\right) \tag{127}$$

严格递增，并有唯一根 $q^\star_\gamma$。另请注意，由（40）中的 $p \in \left(0, \ \frac{1}{2}\right]$ 可知，对所有 $\gamma > 0$，$q_\gamma \le \frac{1}{2}$ 均成立。

我们证明（39）中的下界。由（40）可得

$$\frac{q_\gamma}{1-q_\gamma} = \frac{p^{\frac{1}{\gamma}}}{p^{\frac{1}{\gamma}} + (1-p)^{\frac{1}{\gamma}}} \frac{p^{\frac{1}{\gamma}} + (1-p)^{\frac{1}{\gamma}}}{(1-p)^{\frac{1}{\gamma}}} = \frac{p^{\frac{1}{\gamma}}}{(1-p)^{\frac{1}{\gamma}}}, \tag{128}$$

或等价地

$$\frac{1-p}{p} = \left(\frac{1-q_\gamma}{q_\gamma}\right)^\gamma. \tag{129}$$

将上述结果代入（127），可得

$$D(q_\gamma) = p\,\frac{(1-q_\gamma)^{\gamma-1}}{q_\gamma}\left[\gamma\,\Delta(q_\gamma) + 2q_\gamma - 1\right], \tag{130}$$

其中，对于 $q \le \frac{1}{2}$， $\Delta(q) = (1-q)\log\left(\frac{1}{1-q}\right) - q\log\left(\frac{1}{q}\right) \le 0$。因此，由于 $q_\gamma \le \frac{1}{2}$，我们有

$$D(q_\gamma) \le p\,\frac{(1-q_\gamma)^{\gamma-1}}{q_\gamma}(2q_\gamma - 1) \le 0, \tag{131}$$

这意味着 $q_\gamma \le q_\gamma^\star$。

现在证明（39）中的上界。由（40）可得

$$\frac{1-p}{p} = \left(\frac{1-q_{\gamma+1}}{q_{\gamma+1}}\right)^{\gamma+1}. \tag{132}$$

将上述结果代入（127），得到

$$D(q_{\gamma+1}) = \gamma\, p\,\frac{(1-q_{\gamma+1})^{\gamma-1}}{q_{\gamma+1}^2}\left[(1-q_{\gamma+1})^2\log\left(\tfrac{1}{1-q_{\gamma+1}}\right) - q_{\gamma+1}^2\log\left(\tfrac{1}{q_{\gamma+1}}\right)\right]. \tag{133}$$

对于 $q \le \frac{1}{2}$，我们有 $(1-q)^2\log\left(\frac{1}{1-q}\right) \ge q^2\log\left(\frac{1}{q}\right)$，因此， $D(q_{\gamma+1}) \ge 0$，从而得到 $q_\gamma^\star \le q_{\gamma+1}$。

结合 $D(q_\gamma) \le 0 \le D(q_{\gamma+1})$ 这一事实以及 D 的严格单调性，可得

$$q_\gamma \le q_\gamma^\star \le q_{\gamma+1}. \tag{134}$$

现在证明（41）。根据中值定理，存在 $\xi \in (\gamma,\ \gamma+1)$ 使得

$$q'(\xi) = q_{\gamma+1} - q_\gamma = \widetilde{Q}_{\gamma+1}(1) - \widetilde{Q}_\gamma(1), \tag{135}$$

其中 $q'(\xi) = \frac{\mathrm{d}q_\gamma}{\mathrm{d}\gamma}\Big|_{\gamma=\xi}$，可由（40）计算得出，即

$$\begin{aligned} q'(\xi) &= -\log\left(\frac{p}{1-p}\right)\frac{q_\xi(1-q_\xi)}{\xi^2} && (136)\\ &\le \left|\log\left(\frac{p}{1-p}\right)\right|\frac{1}{4\xi^2} && (137)\\ &\le \left|\log\left(\frac{p}{1-p}\right)\right|\frac{1}{4\gamma^2}, && (138)\end{aligned}$$

其中，（137）成立是因为 $q_\gamma(1-q_\gamma) \le \frac{1}{4}$ for $q_\gamma \le \frac{1}{2}$，而（138）成立是因为 $\xi \in (\gamma,\ \gamma+1)$。结合（135）和（138），可得

$$\widetilde{Q}_{\gamma+1}(1) - \widetilde{Q}_\gamma(1) \le \left|\log\left(\frac{p}{1-p}\right)\right|\frac{1}{4\gamma^2}. \tag{139}$$

命题 9 的证明至此结束。

C 第 4 节的证明

C.1 命题 10 的证明

设 f_0 使得 $(\mathsf{L}_\gamma')^{-1}(f_0) = \frac{1}{|\mathcal{S}|}$。于是，利用泰勒余项定理，我们有

$$\begin{aligned} P_\gamma^\star(x) &= (\mathsf{L}_\gamma')^{-1}\left(-\frac{\alpha_\gamma^\star}{P_X}(x)\right) && (140)\\ &= \frac{1}{|\mathcal{S}|} - \left(\frac{\alpha_\gamma^\star}{P_X(x)} + f_0\right)\frac{\mathrm{d}}{\mathrm{d}t}(\mathsf{L}_\gamma')^{-1}(t)|_{t=f_0} + R_{x,\infty} && (141)\\ &= \frac{1}{|\mathcal{S}|} - \frac{\left(\frac{\alpha_\gamma^\star}{P_X(x)} + f_0\right)}{\mathsf{L}_\gamma''\left((\mathsf{L}_\gamma')^{-1}(f_0)\right)} + R_{x,\infty} && (142)\end{aligned}$$

$$= \frac{1}{|\mathcal{S}|} - \frac{\left(\frac{\alpha_\gamma^\star}{P_X(x)} + \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\right)}{\mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right)} + R_{x,\infty}, \tag{143}$$

其中最后一步源于 $f_0 = \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)$ 这一事实，且余项由下式给出

$$R_{x,\infty} = \frac{\left(\frac{\alpha_\gamma^\star}{P_X(x)} + \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\right)^2}{2} \frac{\mathrm{d}^2}{\mathrm{d}t^2}(\mathsf{L}'_\gamma)^{-1}(t)\Big|_{t=c} \tag{144}$$

$$= -\frac{\left(\frac{\alpha_\gamma^\star}{P_X(x)} + \mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)\right)^2}{2} \frac{\mathsf{L}'''_\gamma(u)}{\left(\mathsf{L}''_\gamma(u)\right)^3}\Bigg|_{u=(\mathsf{L}'_\gamma)^{-1}(c)}, \tag{145}$$

对于某个 $c \in \left(\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right), -\frac{\alpha_\gamma^\star}{P_X}(x)\right)$。结合 (35) 中关于 $\alpha_\gamma^\star$ 的界，不难看出

$$R_{x,\infty} = O\left(\frac{1}{\gamma}\right). \tag{146}$$

将 (143) 两边对 $x \in \mathcal{S}$ 求和，得到

$$\alpha_\gamma^\star \sum_{x\in\mathcal{S}} \frac{1}{P_X(x)} + |\mathcal{S}|\,\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) - \mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right) \sum_{x\in\mathcal{S}} R_{x,\infty} = 0, \tag{147}$$

这意味着

$$\alpha_\gamma^\star = -\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) \mathsf{HM}(P_X) + \frac{\mathsf{HM}(P_X)}{|\mathcal{S}|} \mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right) \sum_{x\in\mathcal{S}} R_{x,\infty} \tag{148}$$

$$= -\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right) \mathsf{HM}(P_X) + O\left(\gamma\left(1 - \frac{1}{|\mathcal{S}|}\right)^\gamma\right), \tag{149}$$

其中最后一步利用了 (146)。

现在，将 (149) 代入 (143)，可得出结论：当 $\gamma \to \infty$ 时，我们有

$$P_\gamma^\star(x) = \frac{1}{|\mathcal{S}|} + \frac{\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)}{\mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right)} \left(\frac{\mathsf{HM}(P_X)}{P_X(x)} - 1\right) + O\left(\frac{1}{\gamma}\right) \tag{150}$$

$$= \frac{1}{|\mathcal{S}|} + O\left(\frac{1}{\gamma}\right), \tag{151}$$

在最后一个等式中，我们用到了 $\mathsf{L}'_\gamma\left(\frac{1}{|\mathcal{S}|}\right)$ $\mathsf{L}''_\gamma\left(\frac{1}{|\mathcal{S}|}\right) = O\left(\frac{1}{\gamma}\right)$。至此，命题 10 的证明完成。

C.2 定理2的证明

我们考虑两种情形：

情形1： 如果 $p_{(i)} \in (0, p_{\gamma,a}]$ 或 $p_{(i)} \in [p_{\gamma,b}, 1)$，那么 $\alpha_\gamma^\star \geq \phi_\gamma(p_{(i)})$。由于对于 $t \in (-\infty, 0)$，$t \mapsto (\mathsf{L}'_\gamma)^{-1}(t)$ 严格递增（见命题 2），因此我们有

$$p_{(i)} \geq (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\alpha_\gamma^\star}{p_{(i)}}\right). \tag{152}$$

根据 (18)，上述不等式右侧恰好为 $p_{(i)}^\star$，因此，$d_i \geq 0$.

情形2： 如果 $p_{(i)} \in (p_{\gamma,a}, p_{\gamma,b})$，那么 $\alpha_\gamma^\star < \phi_\gamma(p_{(i)})$。由于对于 $t \in (-\infty, 0)$，$t \mapsto (\mathsf{L}'_\gamma)^{-1}(t)$ 严格递增（见命题 2），这意味着

$$p_{(i)} < (\mathsf{L}'_\gamma)^{-1}\left(-\frac{\alpha_\gamma^\star}{p_{(i)}}\right). \tag{153}$$

由 (18)，上述不等式的右侧恰为 $p_{(i)}^\star$，因此 $d_i < 0$。定理 2 的证明至此完成。

C.3 命题11的证明

我们着重考虑 $\alpha_\gamma^\star \geq 1$ 的情形（若 $\alpha_\gamma^\star < 1$，则由引理2，结论显然成立）。若 $\alpha_\gamma^\star \geq 1$，则我们需要$\alpha_\gamma^\star = \phi_\gamma(p_{\gamma,b}) \geq 1$。然而，我们现在证明，对所有 $p \in [1/2, 1)$，都有 $\phi_\gamma(p) < 1$，这意味着$p_{\gamma,b} < 1/2$ （以确保 $\phi_\gamma(p_{\gamma,b}) \geq 1$），因此 $p_{(1)} > p_{\gamma,b}$，即 $d_1 \geq 0$。对所有$p \in [1/2, 1)$，$\phi_\gamma(p) < 1$，这是因为：(i) 由命题3，$\phi_\gamma(p)$ 是单峰的，并且满足 $\phi_\gamma(0^+) = 1$ 以及 $\phi_\gamma(1) = 0$；(ii) $\phi_\gamma(1/2) < 1$；(iii) $p \mapsto \phi_\gamma(p)$ 在 $p \in [1/2, 1)$ 上严格递减（由命题3）。因此，$p_{\max} \geq p_{\gamma,b}$，即 $d_1 \geq 0$。 $d_2 < 0$ 这一事实由推论 2 得出。命题11 的证明至此完成。

C.4 命题12的证明

根据命题3，由于$p_{\gamma,a} < 1/3$，我们有

$$\phi_\gamma(p_{\gamma,a}) > \phi_\gamma\left(\frac{1 - p_{\gamma,a}}{2}\right). \tag{154}$$

此外，我们还需要

$$\alpha_\gamma^\star = \phi_\gamma(p_{\gamma,a}) = \phi_\gamma(p_{\gamma,b}). \tag{155}$$

上述两个等式，结合 ϕ_γ 的单峰性（见命题3），意味着

$$p_{\gamma,b} < \frac{1 - p_{\gamma,a}}{2}. \tag{156}$$

现在我们可以证明 $d_1 \geq 0$了。证明采用反证法。假设 $d_1 < 0$。那么，由定理 2 可得

$$p_{(1)} < p_{\gamma,b} < \frac{1 - p_{\gamma,a}}{2}, \tag{157}$$

这将导出

$$\begin{aligned} p_{(1)} + p_{(2)} + p_{(3)} &\leq 2\,p_{(1)} + p_{(3)} && (158)\\ &< 2\frac{1 - p_{\gamma,a}}{2} + p_{\gamma,a} && (159)\\ &= 1, && (160) \end{aligned}$$

其中，第二个不等式成立，是因为在序列 $\{d_i\}$，$i \in \{1, 2, 3\}$ 中至少需要一次变号（参见推论 2）。因此，我们得出了矛盾，因为我们需要 $p_{(1)} + p_{(2)} + p_{(3)} = 1$。至此，命题12的证明完成。

C.5 示例：$p_{\min} < p_{\gamma,\ a}$

设 $\gamma = 0.2$ 和 $\mathcal{S} = \{x_1, x_2, x_3, x_4\}$ 满足

$$P_X(x) = \frac{1}{51}\begin{cases} 1 & x = x_1,\\ 20 & x = x_2, x_4,\\ 10 & x = x_3,\\ 0 & \text{otherwise.} \end{cases} \tag{161}$$

由此，根据式 (18)，我们得到

$$P_\gamma^\star = \begin{cases} 3/154 & x = x_1,\\ 195/499 & x = x_2, x_4,\\ 191/960 & x = x_3,\\ 0 & \text{otherwise.} \end{cases} \tag{162}$$

由此可得

$$d_1 = d_2 = \frac{20}{51} - \frac{195}{499} = \frac{35}{25449} > 0, \tag{163}$$

$$d_3 = \frac{10}{51} - \frac{191}{960} = -\frac{47}{16320} < 0, \tag{164}$$

$$d_4 = \frac{1}{51} - \frac{3}{154} = \frac{1}{7854} > 0, \tag{165}$$

也就是说，$p_{\min} = \frac{1}{51} < p_{\gamma,\ a}$ （另见图7）。

C.6 命题 14 的证明

我们首先注意到，式（36）中的界以及 $\phi(p_{(i)})$ 关于 i 严格递增这一事实（见命题3）意味着存在唯一的 k_M，使得

$$k_M = \max\{i : \phi(p_{(i)}) \le \alpha_\gamma^\star\},\ k_M \in \mathbb{N}. \tag{166}$$

现在，我们考虑两种情形：

1. **情形1：** $i \le k_M$。对于这种情形，我们有 $\phi(p_{(i)}) \le \alpha_\gamma^\star$，这意味着

$$-\frac{\alpha_\gamma^\star}{p_{(i)}} \le -\frac{\phi(p_{(i)})}{p_{(i)}} = \mathsf{L}'_\gamma(p_{(i)}), \tag{167}$$

因此，在上述不等式两边应用 $\left(\mathsf{L}'_\gamma\right)^{-1}$，并回顾 $\left(\mathsf{L}'_\gamma\right)^{-1}$ 在其定义域（$-\infty$，0）上严格递增（见命题 2），可得

$$p^\star_{(i)} \le p_{(i)}, \tag{168}$$

也就是说，$d_i \ge 0$。

2. **情形2：** $i > k_M$。对于这种情形，我们有 $\phi(p_{(i)}) > \alpha_\gamma^\star$，这意味着

$$-\frac{\alpha_\gamma^\star}{p_{(i)}} > -\frac{\phi(p_{(i)})}{p_{(i)}} = \mathsf{L}'_\gamma(p_{(i)}), \tag{169}$$

因此，采用与情形1类似的步骤，我们得到

$$p^\star_{(i)} > p_{(i)}, \tag{170}$$

也就是说，$d_i < 0$。

总之，我们已经证明，对于所有 $i \le k_M$，$d_i \ge 0$；对于所有 $i > k_M$，$d_i < 0$，也就是说，$p_{\min} > p_{\gamma_a}$（另见图 7）。命题 14 的证明到此结束。

注 7. 通过采用与上述类似的步骤，不难证明，只要 $\gamma < \kappa(p_{\max})$，那么 $p_{\max} < p_{\gamma,b}$，也就是说，对于这些 γ 的取值，过度抑制区间确实 存在。

C.7 命题 15 的证明

我们首先指出，若 $\alpha_\gamma^\star < 1$，则结论显然成立，因为根据推论 2，序列 $\{d_i\}$ 至少有一次变号。因此，接下来假设 $\alpha_\gamma^\star \ge 1$。

根据 $\phi_\gamma(p)$ 的定义以及 $\phi_\gamma(p_{\gamma,b}) = \alpha_\gamma^\star$ 这一事实，我们有

$$\alpha_\gamma^\star = p_{\gamma,b}\,(1-p_{\gamma,b})^{\gamma-1}\left(\gamma\log\frac{1}{p_{\gamma,b}} + \frac{1-p_{\gamma,b}}{p_{\gamma,b}}\right) \tag{171}$$

$$\le p_{\gamma,b}\,(1-p_{\gamma,b})^{\gamma-1}\left(\frac{\gamma}{p_{\gamma,b}} - \gamma + \frac{1-p_{\gamma,b}}{p_{\gamma,b}}\right) \tag{172}$$

$$= (1-p_{\gamma,b})^{\gamma}(\gamma+1), \tag{173}$$

其中，不等式成立是因为对于 $x > 0$，$\log(x) \le x - 1$。现在，由于 $\alpha_\gamma^\star \ge 1$，我们可以进一步将上式界定如下，

$$1 \le (1-p_{\gamma,b})^{\gamma}(\gamma+1), \tag{174}$$

这意味着

$$p_{\gamma,b} \le 1 - \left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}}. \tag{175}$$

要证明 $p_{\max} > p_{\gamma,\ b}$，只需证明 $p_{\gamma,\ b} < \frac{1}{|\mathcal{S}|}$ （因为 $p_{\max} \geq \frac{1}{|\mathcal{S}|}$）。将上述上界设为小于 $\frac{1}{|\mathcal{S}|}$，即可得到一个充分条件，即:

$$1 - \left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}} < \frac{1}{|\mathcal{S}|}, \tag{176}$$

或等价地

$$\left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}} > \frac{|\mathcal{S}|-1}{|\mathcal{S}|}. \tag{177}$$

由于 $\gamma \mapsto \left(\frac{1}{\gamma+1}\right)^{\frac{1}{\gamma}}$ 严格递增，只需找到 $\gamma_0 > 0$ 使得

$$\left(\frac{1}{\gamma_0+1}\right)^{\frac{1}{\gamma_0}} = \frac{|\mathcal{S}|-1}{|\mathcal{S}|} \tag{178}$$

以确保 (177) 对所有 $\gamma > \gamma_0$均成立。现在我们着重求解(178)，这等价于

$$\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)^{\gamma_0} = \gamma_0 + 1, \tag{179}$$

or

$$\exp\left(\gamma_0 \log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)\right) = \gamma_0 + 1. \tag{180}$$

令 $x = \gamma_0 + 1$ （其中 $x > 1$）且 $n = \log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)$，得到

$$\exp\left((x-1)\, n\right) = x, \tag{181}$$

或等价地

$$\exp(-n) = x \exp(-nx), \tag{182}$$

或等价地

$$-n \exp(-n) = -n\ x \exp(-nx). \tag{183}$$

令 $y = -n\ x$ 和 $t = -n\exp(-n)$，可得

$$y \exp(y) = t. \tag{184}$$

注意 上述方程只有在 y仅当 $t \geq -\frac{1}{\mathrm{e}}$时才能求解，而在我们的情形下这一条件成立，因为

$$\begin{aligned}
t &= -n \exp(-n) && (185)\\
&= -\log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right) \exp\left(-\log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)\right) && (186)\\
&= \frac{|\mathcal{S}|-1}{|\mathcal{S}|} \log\left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right) && (187)\\
&\geq \frac{1}{2} \log\left(\frac{1}{2}\right) && (188)\\
&> -\frac{1}{\mathrm{e}}, && (189)
\end{aligned}$$

其中，（188）中的不等式成立，是因为（187）中的函数关于 $|\mathcal{S}|$ 单调递增。因此，（184）可以求解；由于 $t < 0$, 它有两个不同的实数解，即

$$y = \mathrm{W}_0(t), \tag{190}$$

and

$$y = \mathrm{W}_{-1}(t). \tag{191}$$

现在我们证明，解 $y = \mathrm{W}_0(t)$ 导致平凡解 $\gamma_0 = 0$。首先注意到，（190）等价于

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_0\left(-n\exp(-n)\right). \tag{192}$$

现在，我们知道当 $-\frac{1}{e} < t < 0$ 时，$\mathrm{W}_0(t) \in (-1, 0)$，因此 $-\frac{1}{n}\mathrm{W}_0(t) \in \left(0, \frac{1}{n}\right)$。于是，由 (192) 可得

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_0\left(-n\exp(-n)\right) \in \left(-1, -1 + \frac{1}{n}\right). \tag{193}$$

注意到 $0 \in \left(-1, -1 + \frac{1}{n}\right)$，且实际上 $\gamma_0 = 0$ 满足 (179)。然而，我们已经知道 $\gamma > 0$。因此，我们只考虑 $y = \mathrm{W}_{-1}(t)$，这等价于

$$\gamma_0 = -1 - \frac{1}{n}\mathrm{W}_{-1}\left(-n\exp(-n)\right), \tag{194}$$

其中回忆一下，$n = \log\left(\frac{|\mathcal{S}|}{|\mathcal{S}|-1}\right)$。还要注意，(194) 的右侧始终为正。这是因为当 $t \le \log\left(\frac{1}{2}\right) = -\log(2)$ 时，$\mathrm{W}_{-1}(t) \in (-\infty, -2\log(2))$，因此 $-\frac{1}{n}\mathrm{W}_{-1}(t) \in \left(\frac{2}{\log(2)n}, \infty\right)$，并且对所有 $|\mathcal{S}| \ge 2$，$\frac{2}{\log(2)n} \ge 2$。因此，我们得出结论：只要 $\gamma > \gamma_0$，就有 $\bar{p}_{\max} > \bar{p}_{\gamma,b}$。至此，命题 15 的证明完成。

C.8 命题 16 的证明

首先注意，当 $p_{\min} > p_{\gamma_a}$ 时，由定理 2 可知，要么 $p_{(i)} \in (p_{\gamma,a}, p_{\gamma,b})$（此时 $d_i < 0$），要么 $p_{(i)} \in [p_{\gamma,b}, 1)$（此时 $d_i \ge 0$），对所有 $i \in \{1, 2, \ldots, |\mathcal{S}|\}$ 均成立。令

$$v = \min\{i : d_i < 0\}. \tag{195}$$

那么，序列 $\{d_i\}$ 有一次变号，即对所有 $i \in \{1, 2, \ldots, v-1\}$，$d_i \ge 0$，且对所有 $i \in \{v, v+1, \ldots, |\mathcal{S}|\}$，$d_i < 0$。此外，我们还知道

$$\sum_{i=1}^{|\mathcal{S}|} d_i = 0. \tag{196}$$

一个分量至多变号一次（从正变负）且分量和为零的向量，其所有部分和均非负；也就是说，对所有 $k \in \{1, \ldots, |\mathcal{S}|\}$，都有

$$\sum_{i=1}^{k} d_i \ge 0, \tag{197}$$

或者等价地

$$\sum_{i=1}^{k} p_{(i)} \ge \sum_{i=1}^{k} p^{\star}_{(i)}, \tag{198}$$

这与 $\sum_{i=1}^{|\mathcal{S}|} p_{(i)} = \sum_{i=1}^{|\mathcal{S}|} p^{\star}_{(i)} = 1$ 一起等价于 $P_X \succ P^{\star}_{\gamma}$。命题 16 的证明至此结束。

C.9 引理3的证明

为证明第一条性质，注意到对于 $p \in (0, 1)$，函数 $\gamma \mapsto p^{(1-p)^{\gamma}}$ 严格递增，因此，$\gamma \mapsto h_{\gamma}(P_X)$ 严格递增。

为证明第二条性质中的下界，注意到

$$h_{\gamma}(P_X) \ge h_0(P_X) = \log\left(\sum_{x \in \mathcal{X}} P_X(x)\right) = 0, \tag{199}$$

其中我们利用了 $\gamma \mapsto h_{\gamma}(P_X)$ 严格递增这一事实。为证明第二条性质中的上界，我们分别考虑两种情形：

- **情形** $\gamma > 1$**：** 在此情形下，我们有

$$\sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma} \tag{200}$$
$$= \sum_{x\in\mathcal{X}} \mathrm{e}^{(1-P_X(x))^\gamma \log P_X(x)} \tag{201}$$
$$\leq \sum_{x\in\mathcal{X}} \mathrm{e}^{(1-\gamma P_X(x)) \log P_X(x)} \tag{202}$$
$$= \sum_{x\in\mathcal{X}} P_X(x)\mathrm{e}^{-\gamma P_X(x) \log P_X(x)} \tag{203}$$
$$\leq \mathrm{e}^{\frac{\gamma}{\mathrm{e}}} \sum_{x\in\mathcal{X}} P_X(x) = \mathrm{e}^{\frac{\gamma}{\mathrm{e}}}, \tag{204}$$

其中，(202) 由伯努利不等式 $(1+x)^r \geq 1+rx, \quad x\geq -1, \quad r\geq 1$ 得出；(204) 则由界 x 对数 $x \geq -^1\mathrm{e}$ 得出。因此，将上述结果代入(57)，我们得到

$$h_\gamma(P_X) \leq \frac{\gamma}{\mathrm{e}}. \tag{205}$$

- **情形** $0 \leq \gamma \leq 1$**：** 对于此情形，我们令

$$f(\gamma) = \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma}, \tag{206}$$

并注意到

$$f'(\gamma) = \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma} (1-P_X(x))^\gamma \log(P_X(x)) \log(1-P_X(x)) \tag{207}$$
$$\leq \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma} \log\left(\frac{1}{P_X(x)}\right) \log\left(\frac{1}{1-P_X(x)}\right) \tag{208}$$
$$\leq \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))^\gamma} \tag{209}$$
$$\leq \sum_{x\in\mathcal{X}} P_X(x)^{(1-P_X(x))} \tag{210}$$
$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}, \tag{211}$$

其中：在 (208) 中，我们用到了这样一个事实：$(1-P_X(x))^\gamma \leq 1$；(209) 由对数$(x) \leq x-1, \quad x>0$ 得出；(210) 是因为 $\gamma \mapsto p^{(1-p)^\gamma}$ 是严格递增的；而在 (211) 中，我们使用了 (204) 中的界。
现在，利用中值定理，存在 $c \in (0, \gamma)$，使得

$$f(\gamma) = f'(c)(\gamma - 0) + f(0) \tag{212}$$
$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma + 1, \tag{213}$$

其中，最后一步用到了 (211) 中的界以及 $f(0) = 1$ 这一事实。接下来，将上述结果代入 (57)，得到

$$h_\gamma(P_X) = \log\left(f(\gamma)\right) \tag{214}$$
$$\leq \log(\mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma + 1) \tag{215}$$
$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma, \tag{216}$$

其中，最后一步用到了界 $\log(1+x) \leq x$.

取两个界中的最大值，即完成引理 3 的证明。

C.10 命题17的证明

为证明第一个表达式，注意到利用式 (55) 中的恒等式可得：

$$H_\gamma(P_X,\ Q_X) \tag{217}$$

$$= \mathbb{E}_{X\sim P_X}\left[\log\left(\frac{1}{Q_X^{(\gamma)}(X)}\right)\right] - h_\gamma(Q_X) \tag{218}$$

$$= \mathbb{E}_{X\sim P_X}\left[\log\left(\frac{P_X(X)}{Q_X^{(\gamma)}(X)}\right)\right] - \mathbb{E}\left[\log P_X(X)\right] - h_\gamma(Q_X). \tag{219}$$

为证明第二个表达式，我们注意到

$$H_\gamma(P_X,\ Q_X) \tag{220}$$

$$= \sum_{x\in\mathcal{X}} P_X(x)(1-Q_X(x))^\gamma \log\left(\frac{1}{Q_X(x)}\right) \tag{221}$$

$$= \rho_\gamma(P_X,Q_X)\left(\sum_{x\in\mathcal{X}} R_X(x)\log\left(\frac{1}{Q_X(x)}\right)\right) \tag{222}$$

$$= \rho_\gamma(P_X,Q_X)\left(D_{\mathsf{KL}}(R_X\|Q_X) + H(R_X)\right). \tag{223}$$

至此，命题17得证。

C.11 命题 18 的证明

下界成立，因为根据命题4，我们知道映射 $\gamma \mapsto H_\gamma(P_X,\ Q_X)$ 是非递增的，因此，

$$\inf_{Q_X} H(P_X,Q_X) - \inf_{Q_X} H_\gamma(P_X,Q_X) \tag{224}$$

$$\geq \inf_{Q_X} H(P_X,Q_X) - \inf_{Q_X} H_0(P_X,Q_X) \tag{225}$$

$$= 0. \tag{226}$$

对于上界，我们有

$$\inf_{Q_X} H(P_X,Q_X) - \inf_{Q_X} H_\gamma(P_X,Q_X) \tag{227}$$

$$= \inf_{Q_X} H(P_X,Q_X) - \inf_{Q_X}\left(H\left(P_X,Q_X^{(\gamma)}\right) - h_\gamma(Q_X)\right) \tag{228}$$

$$\leq \inf_{Q_X} H(P_X,Q_X) - \inf_{Q_X} H\left(P_X,Q_X^{(\gamma)}\right) + \sup_{Q_X} h_\gamma(Q_X) \tag{229}$$

$$\leq \inf_{Q_X} H(P_X,Q_X) - \inf_{Q_X} H(P_X,Q_X) + \sup_{Q_X} h_\gamma(Q_X) \tag{230}$$

$$\overset{\text{上确界}}{=}_{Q_X} h_\gamma(Q_X) \tag{231}$$

$$\leq \mathrm{e}^{\frac{1}{\mathrm{e}}}\gamma, \tag{232}$$

其中：(228) 由 (59) 得出；(229) 是因为 $\inf_x(f(x) - g(x)) \geq \inf_x f(x) - \sup_x g(x)$，这是由于 $f(x) \geq \inf_x f(x)$ 且 $g(x) \leq \sup_x g(x)$；(230) 是因为定义 $Q_X^{(\gamma)}$ 的空间是 $\mathcal{P}(\mathcal{X})$ 的子集；在 (232) 中，我们使用了引理 3 中的界。至此，命题 18 的证明结束。

D 合成数据实验

我们考虑两个类别 $C\in\{0,1\}$，其分别具有 $P[C=0]=0.95$ 和 $P[C=1]=0.05$，以及两个特征 $F_1\in\{0,1,2,3\}$ 和 $F_2\in\{0,1,2,3\}$。请注意，这是一个类别不平衡数据集的示例。我们令

$$P[F_1|C=0] = [0.65, 0.20, 0.10, 0.05], \tag{233}$$

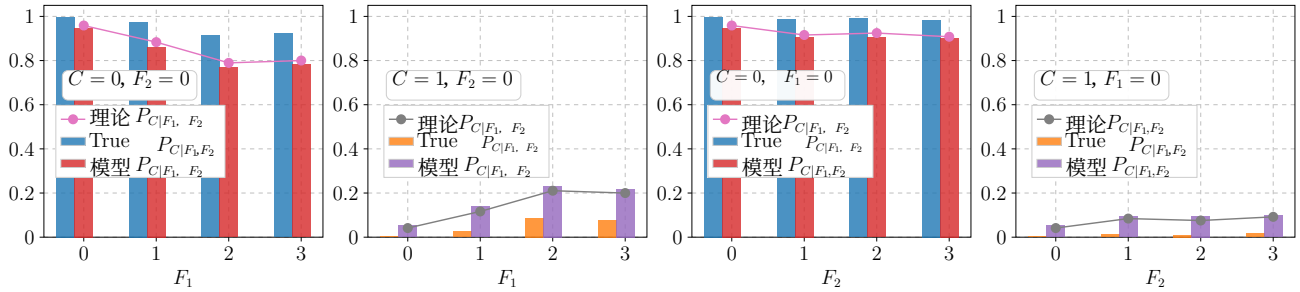


图12：合成数据集。对于每个类别，我们将其中一个特征固定在其众数值，并绘制：真实后验、训练后的神经网络预测的后验概率，以及理论焦点熵极小化器。

$$P[F_1|C=1]=[0.10,0.25,0.45,0.20], \tag{234}$$
$$P[F_2|C=0]=[0.50,0.30,0.15,0.05], \tag{235}$$
$$P[F_2|C=1]=[0.20,0.50,0.20,0.10]. \tag{236}$$

我们假设给定类别后，这两个特征条件独立，即 $P[F_1,F_2|C] = P[F_1|C]\,P[F_2|C]$。因此，我们可以得到如下联合概率质量函数：

$$P[F_1,F_2,C]=P[F_1|C]\,P[F_2|C]\,P[C], \tag{237}$$

并且，根据贝叶斯法则，也得到 $P[C|F_1,F_2]$（一些示例见图12中标注为“True $P_{C|F_1,\ F_2}$”的柱形）。我们将这个 $P_{C|F_1,\ F_2}$ 代入式（18），从而得到 $P_\gamma^\star$对于 $\gamma=1$，即 $P_\gamma^\star=\arg\min_{Q_X} H_\gamma(P_{C|F_1,\ F_2},\ Q_X)$（一些示例见图12中标注为“Theory $P_{C|F_1,\ F_2}$”的柱形）。

我们通过反复从16种可能的二维特征向量空间中采样（即两个特征各有四种可能取值，且有两个类别），生成了10，000个样本。这些样本用于训练一个仅含单个隐藏层的前馈神经网络，该隐藏层有64个神经元，每个神经元均使用 ReLU 激活函数。我们使用式（3）中的焦点损失和 $\gamma=1$ 训练该模型，并采用 Adam 优化器，以 10^{-3}的学习率进行优化。训练时的批次大小为64，共训练30轮。训练完成的神经网络通过对输出层应用 softmax 函数，为两个类别分别生成预测概率（一些示例见图12中标注为“Model $P_{C|F_1,\ F_2}$”的柱形）。从图12可以看出，神经网络输出的概率接近式（18）中的 $P_\gamma^\star$（即最大差异为0.027）。这不仅支持了我们的理论分析，也表明神经网络已有效收敛至全局最小值，在焦点损失下实现了最佳性能。

参考文献

M. M. Al Rahhal、Y. Bazi、H.Almubarak、N. Alajlan和M.AlZuair。结合焦点损失与图像生成的稠密卷积网络用于心电图分类。*IEEE Access*，7:182225–182237，2019年。

Han Bao和Nontawat Charoenphakdee。Calm复合损失：虽为非恰当损失，却是恰当复合损失。载于第*28*届人工智能与统计国际会议，2025年。

Jonas R Brehmer 和 Tilmann Gneiting。适当化：通过贝叶斯行动构造适当评分规则。《统计数学研究所年刊》，72(3):659–673，2020年。A. Buja、W. Stuetzle 和 Y. Shen。用于二分类概率估计和分类的损失函数：结构与应用。工作草稿，*11*月，3:13，2005年。K. Cao、C. Wei、A. Gaidon、N. Arechiga 和 T. Ma。利用标签分布感知边际损失学习不平衡数据集。载于神经信息处理系统进展（*NeurIPS*），第32卷，第1565–1576页，2019年。N. Charoenphakdee、J. Vongkulbhisal、N. Chairatanakul 和 M. Sugiyama。关于用于类别后验概率估计的焦点损失：理论视角。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第5202–5211页，2021年。R. M. Corless、G. H. Gonnet、D. E. G. Hare、D. J. Jeffrey 和 D. E. Knuth。关于兰伯特W函数。《计算数学进展》，5(1):329–359，1996年。

T. M. Cover 和 J. A. Thomas。信息论基础（第*2*版）。Wiley，2006 年。Y. Cui、M. Jia、T.-Y. Lin、Y. Song 和 S. Belongie。基于样本有效数量的类别均衡损失。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第9268–9277页，2019年。J. Dougherty、R. Kohavi 和 M. Sahami。连续特征的有监督和无监督离散化。载于*1995*年机器学习会议论文集，第194–202页。爱思唯尔，1995年。A. Dytso 和 M. Cardone。使用焦点损失的有损信源编码。*arXiv*预印本 *arXiv:2504.19913*，2025年。G. B. Folland。实分析：现代技术及其应用。Wiley，纽约，第2版，1999。A. Ghosh、T. Schaaf 和 M. Gormley。Adafocal：校准感知自适应焦点损失。神经信息处理系统进展（*NeurIPS*），35:1583–1595，2022年。H. He 和 E. A. Garcia。从不平衡数据中学习。*IEEE*知识与数据工程汇刊，21(9):1263–1284，2009年。D. Impedovo 和 G. Pirlo。手写字符识别的分区方法：综述。模式识别，47(3):969–981，2014年。M Jeong、M. Cardone 和 A. Dytso。揭示多类别分类的最优性能。载于第三十七届神经信息处理系统大会，2023年。URL https://openreview.net/ forum?id=p9k5MS0JAL。T. O. Kv˚alseth。熵及其凹性和舒尔凹性条件。*IEEE Access*，10:96006–96015，2022年。Z. Leng、M. Tan、C. Liu、E. D. Cubuk、X. Shi、S. Cheng 和 D. Anguelov。PolyLoss：分类损失函数的多项式展开视角。*arXiv*预印本 *arXiv:2204.12511*，2022年。X. Li、X. Sun、Y. Meng、J. Liang、F. Wu 和 J. Li。面向数据不平衡的自然语言处理任务的Dice损失。载于第*58*届计算语言学协会年会论文集，第4658–4669页，2020a。X. Li、W. Wang、L. Wu、S. Chen、X. Hu、J. Li、J. Tang 和 J. Yang。广义焦点损失：为密集目标检测学习高质量且分布式的边界框。神经信息处理系统进展（*NeurIPS*），33: 21002–21012，2020b。X. Li、C. Lv、W. Wang、G. Li、L. Yang 和 J. Yang。广义焦点损失：面向密集目标检测的高效表征学习。*IEEE*模式分析与机器智能汇刊，45(3):3139–3153，2022年。J. Lin、X. Jin 和 C. Zhang。TCE：用于分类的倾斜交叉熵损失。载于*IEEE/CVF*计算机视觉与模式识别研讨会论文集（*CVPRW*），第4061–4070页，2021年。T.-Y. Lin、P. Goyal、R. Girshick、K. He 和 P. Doll´ar。密集目标检测的焦点损失。载于*IEEE*国际计算机视觉会议论文集（*ICCV*），第2980–2988页，2017年。J. Mukhoti、V. Kulharia、A. Sanyal、S. Golodetz、P. H. S. Torr 和 P. Dokania。使用焦点损失校准深度神经网络。神经信息处理系统进展，33:15288–15299，2020年。B. Peters、V. Niculae 和 A. F. T. Martins。稀疏序列到序列模型。载于第*57*届计算语言学协会年会论文集（*ACL*），第1504–1519页，2019年。M. Ren、W. Yu、J. Ma、S. Zhao、H. Li 和 R. Urtasun。面向长尾视觉识别的均衡元Softmax。载于神经信息处理系统进展（*NeurIPS*），第33卷，第4175–4186页，2020年。W. Shu、S. Wang、Q. Chen、Y. Hu、Z. Cai 和 R. Lin。基于残差网络和焦点损失的乳腺癌病理图像分类。载于*2019*年第三届国际计算机科学与人工智能大会论文集，第211–214页，2019年。J. Vongkulbhisal、P. Vinayavekhin 和 M. Visentini-Scarzanella。通过蒸馏统一异构分类器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第3175–3184页，2019年。Y. Zhang、C. Chen、H. Shi、M. Song、Y. Wu 和 J. Zhang。用于校准分类的双重焦点损失。载于第*40*届国际机器学习大会论文集（*ICML*），2023年。