神经网络校准的期望一致性

Lucas Clart´e[1], BrunoLoureiro[2],
Florent Krzakala[3], 和 Lenka Zdeborov´a[1]

[1] ´洛桑联邦理工学院（EPFL），计算统计物理实验室，CH-1015 洛桑，瑞士[2] D´计算机系，´巴黎高等师范学院 - PSL 与法国国家科学研究中心，45 rue d'Ulm, F-75230 Paris cedex 05, France[3] ´洛桑联邦理工学院（EPFL），信息、学习与物理实验室，CH-1015 洛桑，瑞士

arXiv:2303.02644v2 [cs.LG] 5 Aug 2023

2023年8月8日

摘要

尽管深度神经网络表现惊人，但众所周知，它们往往对自己的预测置信度过于乐观。因此，寻找有效且高效的神经网络校准方法，是提升深度学习不确定性量化能力的重要工作。本文介绍了一种名为期望一致性（EC）的新型校准技术：通过强制验证集平均置信度与正确标签的平均占比相一致，对最后一层权重进行训练后重新缩放。首先，我们表明，在不同神经网络架构和数据集上，EC 方法的校准表现与温度缩放（TS）相近，同时所需的验证样本和计算资源也相当。不过，我们认为，EC 提供了一种有理论依据的方法，其根植于贝叶斯最优性原理——即西森恒等式。接下来，我们对合成环境中的 TS 和 EC 进行渐近刻画，并表明二者的表现关键取决于目标函数。具体而言，我们讨论了 EC 显著优于 TS 的实例。

1 引言

随着深度学习模型日益广泛地应用于人类社会的各个领域，开发可靠的方法以恰当评估其预测结果的可信度变得愈发必要。事实上，人们提出了多种不确定性量化方法，用于衡量训练后的神经网络预测所关联的置信度 [Abdar 等人，2021年，Gawlikowski 等人，2021]。尽管这些指标在实践中广受欢迎，但众所周知，其中一些指标（例如将最后一层 Softmax 分数解读为置信度分数）会导致对真实类别概率的高估 [Guo 等人，2017]。因此，人们提出了各种神经网络校准方法 [Gal 和 Ghahramani，2016，Guo 等人，2017年，Maddox 等，2019年，Minderer 等人，2021]。

在这项工作中，我们提出了一种用于神经网络训练后校准的新方法，称为期望一致性（EC）。该方法通过调整最后一层权重的尺度，使平均置信度与验证集上的平均分类准确率一致。这一过程受到贝叶斯推断文献中最优性条件的启发。因此，除了实现简单且计算高效之外，它还为温度缩放等类似校准技术提供了一种具有数学原理支撑的替代方案。本工作的目标是介绍期望一致性校准方法，展示其在不同深度学习任务中的表现，并在受控情境下提供理论保证。更具体地说，我们的**主要贡献**如下：

• 我们提出了一种新方法期望一致性（EC），用于校准神经网络的训练后预测。该方法通过重新缩放最后一层权重，使平均置信度与验证集上的平均准确率相匹配。我们从贝叶斯推断的角度阐释期望一致性，为其提供数学依据。

• 虽然不确定性量化领域中有许多校准方法，但我们将期望一致性与深度学习实践中一种密切相关且广泛使用的方法进行比较：温度缩放（TS）。我们在不同网络架构和真实数据集上进行的实验表明，这两种方法在实践中得出的结果非常相似。

• 我们在一个存在过度自信问题的高维逻辑回归模型中，对期望一致性进行了理论分析；这类问题与深度神经网络中的问题类似。我们表明，在这种环境下，期望一致性在不同的不确定性指标上始终优于温度缩放。理论分析还阐明了这两种方法结果相似的原因。

本文结构如下：回顾相关文献并介绍期望一致性方法（第3节）后，我们将期望一致性与温度缩放在真实数据上的表现进行比较（第4节），并说明这两种方法的表现相似。作为第4节的补充，我们在第5节对这两种方法进行理论分析，并描述一种期望一致性优于温度缩放的合成环境。

本项目使用的代码可在以下代码库中获取：https://github.com/SPOC-group/expectation-consistency

---

算法1： 期望一致性（EC）

输入： 验证集 $(\boldsymbol{x}_i, y_i)_{i=1}^{n_{val}}$，分类器 $\hat{f}$：$\mathcal{X} \to \mathbb{R}^K$计算逻辑值 $\boldsymbol{z}_i = \hat{f}(\boldsymbol{x}_i) \in \mathbb{R}^K$ 并输出 $\hat{y}_i = \arg\max_k \boldsymbol{z}_{ik}$计算验证集上的准确率 $\mathcal{A}_{val} = \frac{1}{n_{val}} \sum_i \delta(y_i = \hat{y}_i)$ 确定 T_{EC} 使得 $\frac{1}{n_{val}} \sum_i \max_k \sigma^{(k)}(\boldsymbol{z}_i/T) = \mathcal{A}_{val}$**输出：** 温度 T_{EC}，以及新样本上的概率 $\max_k \sigma^{(k)}(\boldsymbol{z}^{\mathrm{new}}/T_{\mathrm{EC}})$，

1.1 相关工作

神经网络的校准 — 预测模型，尤其是神经网络的校准，已得到广泛研究，参见Abdar 等人 [2021], Gawlikowski 等人 [2021]的两篇综述。具体而言，人们发现现代神经网络架构会给出过度自信的预测 [Guo 等人，2017 年，Minderer 等人，2021]。尽管过度自信可能部分归因于过参数化，但一些理论研究 [Bai等人，2021年，Clart´e 等人，2022b，a] 表明，即使是欠参数化区域中的简单回归模型，也可能表现出过度自信。

存在一系列方法，能够在不对数据分布做任何假设的情况下保证渐近校准（即当样本数量足够大时），例如参见 Gupta 等人 [2020]。然而，对于有限的样本数量，目前尚不清楚所提出的方法中哪一种能够提供最准确的校准。

温度缩放 — Guo 等人 [2017] 提出了 温度缩放 (TS)，这是一种简单的后处理方法，通过重新缩放最后一层权重的范数并进行交叉验证来实现。与 Platt 缩放 [Platt，2000] 或直方图分箱 [Zadrozny 和 Elkan相比，2001], TS 简单高效，因此在实践中被广泛用于校准神经网络的输出 [Abdar 等人，2021]。此外，Clart´e 等人 [2022曾] 表明，在某些设定下，就不确定性量化而言，TS 能与成本高得多的贝叶斯方法相媲美。尽管 Gupta 等人 [2020] 表明，在对数据模型不作任何假设的情况下，TS 等单射校准方法通常无法实现校准，Guo 等人 [2017] 得出结论：“温度缩放是最简单、最快捷、最直接的方法，而且令人惊讶的是，它往往也是最有效的方法。”这说明了 TS 为何在实践中得到如此广泛的应用。

贝叶斯方法 — 贝叶斯方法（例如高斯过程）能够在样本数量有限的情况下，在（至少是隐含的）数据分布假设下，开箱即用地估计不确定性。当数据生成过程已知时，估计模型不确定性的最佳方法是使用预测后验。然而，贝叶斯推断通常难以处理，因此人们将若干近似贝叶斯方法应用于神经网络，例如深度集成 [Lakshminarayanan 等，2017年a]或权重平均 [Maddox 等， 2019]。另一方面，温度缩放这类后处理方法的优势在于，它直接作用于网络的未归一化输出，而且无需额外训练。Kristiadi 等人提出了一种类似的贝叶斯方法 [2020], ，其中对最后一层权重采用高斯分布。贝叶斯方法通常涉及从高维后验中采样 [Mattei， 2019], 并且已有多种方法被提出，以高效地计算这些采样结果[Graves，2011年，Gal 和 Ghahramani，2016，Lakshminarayanan 等，2017b，Maddox 等，2019]。

记号——我们记 $[n] := \{1, \cdots, n\}$；$\mathbf{1}(A)$ 表示集合 A 的指示函数； $\mathcal{N}(\boldsymbol{x}|\mu, \Sigma)$ 表示均值为 μ 、协方差为 Σ 的多元高斯分布概率密度函数。

2 设置

考虑一个 K 类分类问题，其中神经网络分类器在数据集 $(\boldsymbol{x}_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times [K]$ 上进行训练。不失一般性，对于给定输入 $\boldsymbol{x} \in \mathbb{R}^d$，我们可以将分类器的输出写成一个 K 维向量 $\boldsymbol{z}(\boldsymbol{x}) = W\boldsymbol{\varphi}(\boldsymbol{x}) \in \mathbb{R}^K$，其中我们将最后一层特征记为 $\boldsymbol{\varphi}: \mathbb{R}^d \to \mathbb{R}^p$，将输出层权重记为 $W \in \mathbb{R}^{K\times p}$。我们将类别 k 的预测置信度定义为：

$$\hat{f}(\boldsymbol{x}, k) := \sigma^{(k)}(\boldsymbol{z}(\boldsymbol{x})) = \frac{e^{z_k(\boldsymbol{x})}}{\sum_{l\in[K]} e^{z_l(\boldsymbol{x})}} \in (0, 1) \tag{1}$$

其中 $\sigma: \mathbb{R}^K \to (0, 1)^K$ 是 softmax 激活函数。简而言之，$\hat{}\, f(\boldsymbol{x}, k)$ 定义了网络估计的一个概率，即 $\boldsymbol{x}$ 属于类别 k 的概率。对于给定的 $\boldsymbol{x} \in \mathbb{R}^d$，模型的最终预测为 $\hat{}\,\hat{}\,\hat{}\, y(\boldsymbol{x}) = \arg\max_k f(\boldsymbol{x}, k) \in [K]$，相应的预测置信度 $f(\boldsymbol{x}) = \max_k f(\boldsymbol{x}, k) = \hat{}\, f(\boldsymbol{x}, \hat{}\, y(\boldsymbol{x})) \in (0, 1)$。按照惯例，以下我们主要关注这样一种情况：网络通过使用交叉熵损失最小化经验风险（ERM）进行训练：

$$\begin{aligned}\ell(\hat{f}(\boldsymbol{x}), y) &= -\log \hat{f}(\boldsymbol{x}, y) \\ &= \sum_{k=1}^{K} \delta(y = k) \log \sigma^{(k)}(W\boldsymbol{\phi}(\boldsymbol{x})),\end{aligned}$$

尽管这里介绍的许多概念可以直接推广到其他训练过程。训练质量通常通过模型在未见数据上的泛化能力来评估。这可以通过测试误分类误差和测试损失函数来量化：

$$\mathcal{E}_g = \mathbb{E}_{\boldsymbol{x},y}\left[\delta\left(\hat{y}(\boldsymbol{x}) \neq y\right)\right], \quad \mathcal{L}_g = -\mathbb{E}_{\boldsymbol{x},y}\left[\log \hat{f}(\boldsymbol{x}, y)\right]$$

这些是点性能指标。然而，我们通常也希望量化网络预测置信度的质量。文献中提出了多种不确定性指标，但目前较新的指标包括校准、期望校准误差（ECE）和布里尔分数（BS），定义如下：

$$\begin{cases}\Delta_p &= p - \mathbb{P}_{\boldsymbol{x},y}\left(\hat{y}(\boldsymbol{x}) = y | \hat{f}(\boldsymbol{x}) = p\right) \\ \text{ECE} &= \mathbb{E}_{\boldsymbol{x},y}\left(|\Delta_{\hat{f}(\boldsymbol{x})}|\right) \\ BS &= \mathbb{E}_{\boldsymbol{x},y}\left(\sum_{k=1}^{K}(\hat{f}(\boldsymbol{x}, k) - \delta(y = k))^2\right)\end{cases} \tag{2}$$

数据集	模型	$\mathcal{E}_g$	T_{TS}	T_{EC}	ECE	ECE_{TS}	ECE_{EC}	BS	BS_{TS}	BS_{EC}
SVHN	ResNet-20	6.8 %	1.59	1.55	2.6 %	1.5 %	1.3 %	10.5 %	10.4 %	10.4 %
CIFAR10	ResNet-20	13.5 %	1.37	1.38	5.3 %	1.9 %	1.9 %	20.0 %	19.43%	19.2 %
CIFAR10	ResNet-56	13.1 %	1.42	1.43	6.0%	2.5 %	2.4 %	20.2 %	19.3 %	19.3 %
CIFAR10	DenseNet-121	12.5 %	1.78	1.86	7.9 %	3.0 %	2.5 %	20.4 %	18.6 %	18.5 %
CIFAR100	ResNet-20	31.0%	1.44	1.44	10.2%	1.7%	1.7%	44.3 %	42.5 %	42.5 %
CIFAR100	ResNet-56	27.3 %	1.73	1.79	14 %	2.6 %	2.2 %	41 %	38.0 %	38.0 %
CIFAR100	VGG19	26.4 %	2.14	2.28	19.9 %	5.3 %	4.8 %	44.8 %	37.2 %	36.9 %
CIFAR100	RepVGG-A2	22.5 %	1.07	1.16	5.3 %	4.6 %	4.4 %	32.1 %	31.9 %	32.0 %

图 1：比较温度缩放（TS）和期望一致性（EC）在各种模型和数据集上的期望校准误差（ECE）与 Brier 分数（BS）。我们发现这两种校准方法之间的差异非常小。鉴于 TS 在实践中的表现如此出色，我们推测 EC 至少也能达到同样的效果。

ˆ 注意，布里尔分数是一种适当损失，这意味着当 $f(\boldsymbol{x},\ k)$ 是真实的边缘分布 ˆ $\mathbb{P}(y=k|\boldsymbol{x})$ 时，它会达到最小值。ECE 则并非如此：事实上，定义为 $f(\boldsymbol{x},\ k)=\mathbb{P}(y=k)$ 的估计量虽然 ECE 为 0，但并不对应于给定 $\boldsymbol{x}$ 条件下 y 的边缘分布，而且测试误差并非最优。最后，我们引入带温度参数 $T>0$ 的置信度函数：

$$\hat{f}_T(\boldsymbol{x},k)=\sigma^{(k)}(W\varphi(\boldsymbol{x})/T). \tag{3}$$

3 期望一致性校准

本文提出的方法与温度缩放方法类似，Guo 等人 [2017] 在相关工作部分讨论过该方法，但在温度参数的选取方式上存在关键差异。下文还将以广受欢迎且应用广泛的温度缩放（TS）流程作为主要基准。

温度缩放——尽管式（1）中引入的基于分数的置信度度量看起来可能很自然，但数值证据表明，对于现代神经网络架构，它往往会过度自信[Guo 等人，2017]。换言之，它会高估类别归属的概率。为缓解过度自信问题，Guo 等人 [2017] 提出了一种称为温度缩放（TS）的后训练校准方法[Minderer 等人, 2021年, Wang 等人，2021]。温度缩放通过一个正数常量 $T>0$ （即“温度”）对训练好的网络输出$\boldsymbol{z}\mapsto \boldsymbol{z}/T$ 重新缩放，再通过调整该常量来校准预测置信度。等价地，TS 可视为对最后一层权重的范数进行重新缩放 W。Guo 等人[2017] 发现，选择使验证集上的交叉熵损失最小的 T {$(\boldsymbol{x}_i,y_i)^{i\in[n_{val}]}$:

$$T_{\mathrm{TS}}=\arg\min_{T>0}\left(-\sum_{i=1}^{n_{val}}\ell(\hat{f}_T(\boldsymbol{x}_i,y_i))\right) \tag{4}$$

ˆ 会得到校准效果更好的重缩放预测器 $f_{T_{\mathrm{TS}}}$。为了直观了解其对置信度的影响，不妨考察 TS 的两个极端情况。一方面，如果 $T\ll 1$，Softmax 将由置信度最大的类别主导，最终收敛为硬阈值化 $T\to 0^{+}$。这通常会导致预测器过度自信。另一方面，对于 $T\gg 1$，Softmax 对训练得到的权重会越来越不敏感，并在 $T\to\infty$ 处收敛为均匀向量。这通常对应于一个置信度不足的预测器。因此，通过调整 T，我们既可以让预测器降低过度自信程度（通过降低温度 $T<1$），也可以减轻置信度不足的问题（通过提高温度 $T>1$）。

温度缩放是矩阵/向量缩放的一种具体形式，其中逻辑值 z_i 在经过 Softmax 处理之前，会先乘以一个矩阵/向量。尽管矩阵缩放和向量缩放更为通用，但 Guo 等人Guo 等人 [2017]发现它们的表现不如温度缩放（TS）。人们开发了温度缩放的不同变体。与向量缩放类似，

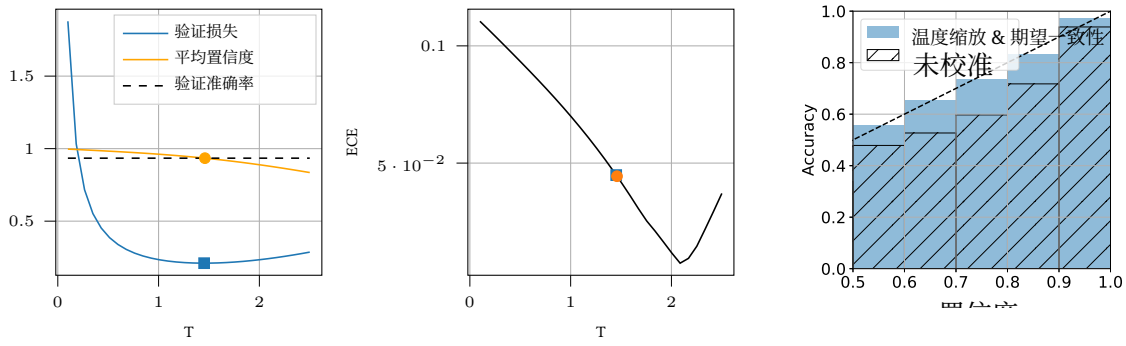


图2：左：模型的验证损失和平均置信度随温度 T 的变化情况；该模型为在 CIFAR10 上训练的 DenseNet121。深色虚线表示验证集的准确率。橙色（蓝色）叉号分别对应 T_{EC}, T_{TS}。中：模型的 ECE 随 T 的变化情况，蓝色和橙色圆点分别对应 TS 和 EC。右：在 CIFAR10 上训练的 Resnet20 在温度缩放前后的可靠性图。EC 后的可靠性图与 TS 后的可靠性图无法区分。

基于类别的温度缩放 [Frenkel 和 Goldberger， 2021] 为每个类别计算一个温度，并通过最小化验证集上的期望校准误差（ECE），而非验证损失函数，来确定最佳温度。虽然温度缩放可以自然地应用于神经网络的最后一层输出，Kull 等人 [2019] 将温度缩放扩展到了更通用的多分类模型。

期望一致性 — 在这项工作中，我们提出了一种新颖的校准方法，以下简称ˆ为期望一致性 (EC)。与 TS 一样，起点是一个预训练置信度函数 f，我们ˆ通过引入温度进行缩放 f_T $T > 0$。关键区别在于我们调整温度的过程。我们不最小化验证损失（4），而是寻找一个温度，使平均置信度等于测试集中正确标签的比例。用数学术语来说，我们定义 T_{EC}使其满足以下条件：

$$\frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \hat{f}_{T_{\mathrm{EC}}}(\boldsymbol{x}_i) = \frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \mathbf{1}(\hat{y}(\boldsymbol{x}_i) = y_i) \tag{5}$$

这一选择背后的直觉如下：校准分类器应满足，对于所有 $p \in (0,\ 1)$ $\Delta_p = 0$，都成立。通过调整温度参数 T无法实现这一条件，因此，一个不那么严格的条件是要求其在期望意义下成立 $\mathbb{E}_{\boldsymbol{x}}\left[\Delta_{\hat{f}(\boldsymbol{x})}\right] = 0$，从而确保分类器在平均意义上经过校准。这等价于要求平均置信度等于在验证集上预测正确类别的概率。请注意，我们直接将其与验证集上的置信度进行比较，这类似于保形预测中所采用的 [Papadopoulos 等人， 2002] 用于估计预测集的方法（而不是我们在此旨在实现的校准）。

从贝叶斯视角考察 EC 很有启发性。就本段而言，假设训练数据和验证数据均独立地从参数化概率分布$p(\boldsymbol{x}, y|\theta)$中抽取。如果我们能够获得数据的分布（但无法得知参数θ的具体取值），那么贝叶斯最优置信度函数将是 $f_\star(\boldsymbol{x}|\theta) = p(y|\boldsymbol{x}, \theta)$ 关于给定训练数据时权重的后验分布 $p(\theta|(\boldsymbol{x}_i, y_i)_{i\in[n]})$ 的期望。在这种情况下，甚至不需要验证集，因为后验分布下的不确定性就能预测期望测试准确率。在第5.1节中，我们将用一个具体的数据分布来说明这一讨论。贝叶斯最优预测器的这种期望一致性性质，在信息论和统计物理学文献中称为 Nishimori 条件 [Iba，1999，M´easson 等人，2009，Zdeborov´a 和 Krza- kala， 2016]。因此，从这个视角来看，要求条件 (5) 成立，可以理解为对重新缩放后的置信度函数施加 Nishimori 条件。Nishimori 条件也用于期望最大化算法中，以学习超参数 Dempster 等人 [1977]。我们将在第5节介绍如何将温度缩放和期望一致性都理解为学习过程

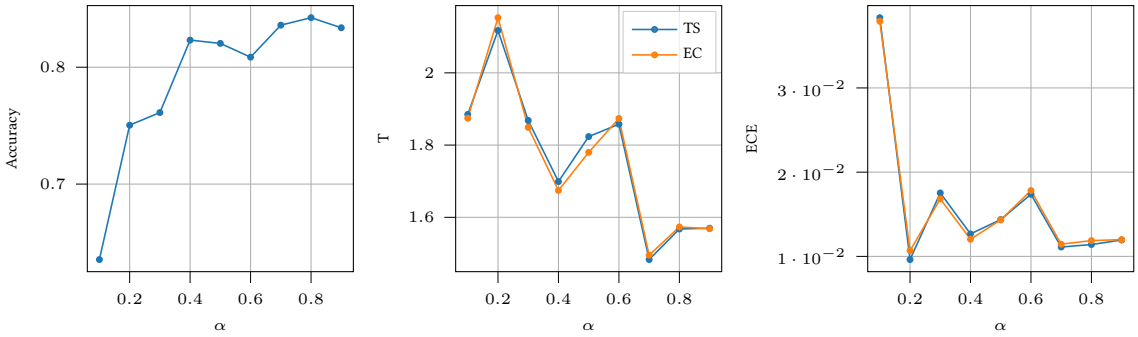


图3：左：ResNet-20 模型的准确率（左）、温度缩放和期望一致性返回的温度（中），以及模型的期望校准误差（右）随训练集大小的变化 $\alpha = n_{\text{train}}/50000$。模型使用与图1中相同的超参数进行训练。再次可以看到，即使样本量差异很大，这两种方法的表现仍然相当。

.

针对超参数 T。此处提出的 EC 方法的核心思想是，即使不知道数据生成模型，经过校准的不确定性量化方法也应满足期望一致性（5）关系。

请注意， T_{EC} 存在且唯一。事实上，平均置信度是温度的递减函数；当 $T \to 0^+$ 时，它趋近于一；当 $T \to \infty$时，它趋近于零。因此，存在唯一的 T_{EC} 满足约束（5），实际中可以通过二分法求得。图2 展示了 T_{EC}的唯一性。此外，请注意，期望一致性比温度缩放更加灵活：在多分类问题中，对于任意 $N \in [K]$，我们都可以固定温度，使平均置信度等于前 N 准确率。本文关注前1 准确率。

4 在真实数据上的实验

本节将介绍在真实数据集上开展的数值实验，并比较 EC 和 TS 的性能。正如我们将看到的，在实际场景中，两种方法都能取得相近的校准性能。

实验设置 — 我们研究第 3 节中的校准方法在图像分类任务中的性能。实验在三个常用的图像分类数据集上进行：

• SVHN Netzer 等人 [2011] 由彩色 32×32 标注数字图像组成。训练集/验证集/测试集的大小分别为 65931/7325/26032。

• CIFAR10 和 CIFAR100 数据集 Krizhevsky [2009], 包含 32×32 来自 10/100 个类别（分别为狗、猫、飞机等）的彩色图像。CIFAR10 的训练集/验证集/测试集分别包含 45000/5000/10000 张图像，CIFAR100 分别包含 50000/5000/5000 张图像。

我们研究了针对图像分类任务调整的不同神经网络架构：残差网络 [He 等人，2016], 密集连接网络 [Huang 等人， 2017], VGG [Simonyan 和 Zisserman， 2014] 以及 RepVGG [Ding 等人， 2021]。对于 CIFAR100，我们采用了在线提供的预训练模型。有关训练过程的更多细节，请参见附录 A。

结果——我们参考表1，比较上述不同数据集和模型上的TS与EC。有趣的是，我们观察到，在不同任务和架构中，EC和TS得出的温度都非常相近，这意味着二者的ECE和布里尔分数也相近。特别值得注意的是，两种方法都给出 $T > 1$，这与原始网络过度自信这一事实相符。因此，正如预期的那样，两种方法都改善了分类器的校准。

图 2 右侧面板展示了在 CIFAR10 上训练的 ResNet 20 的可靠性图：我们观察到，在应用温度缩放和期望一致性之前，准确率低于置信度。换言之，模型过度自信，而温度缩放和期望一致性都能改善模型的校准。

请注意，这两种方法都能改善布里尔分数，且结果非常相似。从计算成本来看，期望一致性的运行效率与温度缩放相当，只需几行代码即可实现，详见 GitHub代码仓库，我们在其中提供了用于复现本文所讨论实验的代码。不过，我们认为期望一致性是一种理论依据更充分的校准方法，因为它将模型的置信度约束为与准确率相对应，并且具有自然的贝叶斯解释。此外，正如我们将在第 5 节中讨论的那样，我们可以为期望一致性推导出明确的理论结果。

我们的实验表明，TS 与 EC 之间的相似性不受模型准确率的影响。事实上，在图3中，我们展示了使用不同数量的数据训练的残差网络模型的准确率和 ECE。正如预期，模型的准确率会随着训练数据量的增加而提高。从中间和右侧面板可以看出，无论模型的准确率如何，两种方法得到的温度和 ECE 都极为相似。最后，我们在图2的中间面板中绘制了 ECE 随温度变化的曲线，并观察到，T_{TS} 和 T_{EC} 都没有接近 ECE 的最小值。不过，正如我们在第3节所讨论的，ECE 只是一个不确定性量化指标，并非适当损失，因此我们不希望专门针对这一指标来优化温度。

受损数据上的实验 — 在附录C中，我们比较了 EC 和 TS 在一项图像分类任务上的表现，其中测试数据遭到了破坏。我们使用与第4节相同的数据集和架构，但随机选取某些图像类别的测试数据目标标签。引入类别相关噪声的目的是评估两种方法在更现实的场景中的表现；在这种场景下，训练数据与测试数据之间存在分布偏移，相关做法见?。我们在附录的表1中报告了 EC 和 TS 在 ECE 和布里尔分数方面的表现。我们观察到，EC 平均可使测试 ECE 降低 7%，这表明在更现实的场景中，EC 是 TS 的一种有竞争力的替代方案。请注意，在这种设置下，EC 和 TS 得到的温度不同，这与表1中描述的结果相反。完整的实验细节见附录C。

5 期望一致性的理论分析

正如第 4 节所示，我们在真实数据和神经网络模型上进行的实验表明，尽管 EC 和温度缩放的性质不同，但它们在不同架构和数据集上的校准表现相似。本节将研究特定设定下的 EC 和温度缩放，在这些设定中，我们可以推导出关于其校准性质的理论保证。

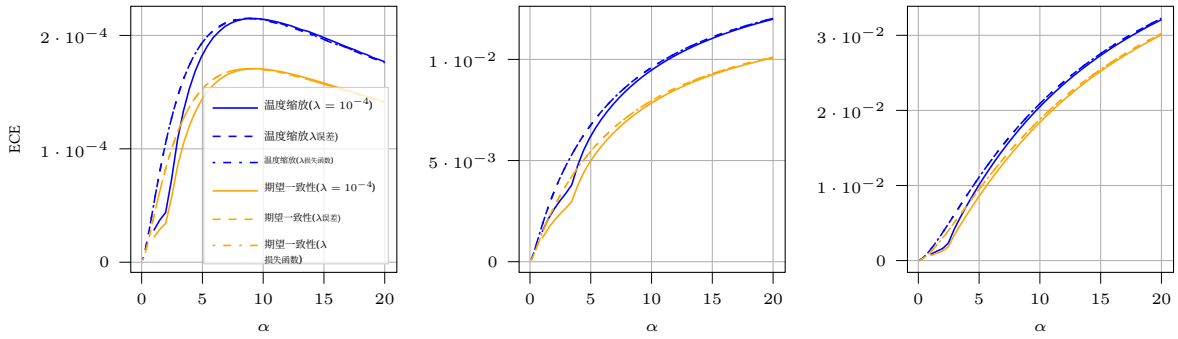


图4：具有三个不同 λ 值（10^{-4}、 λ_{error}、 λ_{loss}）的正则化逻辑回归的期望校准误差：未校准、经过温度缩放后以及经过期望一致性处理后。从左到右分别为： $\sigma_\star$ $=\sigma_{\text{logit}}$、 σ_{affine}、 σ_{constant} 。

为便于具体说明，在接下来的示例中，我们将重点讨论二元分类问题。对此，不失一般性，我们可以假设 $y \in \{-1, +1\}$。在这种编码方式下，Softmax 函数等价于逻辑函数 $\sigma(t) := (1 + e^{-t})^{-1}$，而硬最大值由符号函数给出。此外，我们假设训练集 $(\boldsymbol{x}_i,\ y_i)_{\ i\in[n]}$ 和验证集 $(\boldsymbol{x}_i,\ y_i)_{\ i\in[n_{val}]}$ 均独立地从以下数据中抽取

生成模型：

$$f_\star(\boldsymbol{x}) \coloneqq \mathbb{P}(y^\mu = 1|\boldsymbol{x}^\mu) = \sigma_\star\left(\frac{\boldsymbol{w}_*^\top \boldsymbol{x}^\mu}{T_\star}\right)$$
$$\boldsymbol{x}^\mu \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_d \mathbf{I}_d), \quad \boldsymbol{w}_* \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d) \tag{6}$$

带有 $\sigma_\star : \mathbb{R} \to (0, 1)$ 个激活函数，并且 $T_* > 0$ 显式地对权重范数进行参数化 ts.

首先，在第 5.1 节中，我们通过一个 $T_{\mathrm{TS}} = T_{\mathrm{EC}} = T_\star$ 的例子，对 TS 和 EC 方法进行贝叶斯解释。接着，在第 5.2 节中，我们分析一个失配经验风险最小化的情形，在这种情况下，两种方法会得出不同的结果。最后，在第 5.3 节中，我们讨论 EC 始终优于 TS 的一种情况。

5.1 与贝叶斯估计的关系

考虑一个贝叶斯推断问题：给定训练数据 $\mathcal{D} : = \{(\boldsymbol{x}_i, \ y_i)_{i\in[n]}\}$，什么样的预测器能够使准确率最大化？如果统计学家能够完全了解数据生成过程（6），那么只需对给定数据时权重的后验分布下的数据似然进行积分即可得到该预测器：

$$f_{\mathrm{bo}}(\boldsymbol{x}) \coloneqq \mathbb{P}(y = 1|\mathcal{D}, \boldsymbol{x}) = \int \mathrm{d}\boldsymbol{w}\ \sigma\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{T_*}\right) p(\boldsymbol{w}|\mathcal{D}, T_\star)$$

其中，后验分布明确表示为：

$$p(\boldsymbol{w}|\mathcal{D}, T_\star) \propto \mathcal{N}(\boldsymbol{w}|0, \mathbf{I}_d) \prod_{i\in[n]} \sigma_*\left(y_i \frac{\boldsymbol{w}^\top \boldsymbol{x}_i}{T_\star}\right). \tag{7}$$

上述贝叶斯最优预测器经过了良好校准 [Clart′等人，2022b], 因此满足期望一致性条件：其平均置信度等于准确率。现在考虑这样一种情形：统计学家仅掌握有关数据生成过程的部分信息：她知道先验分布和似然，但无法获得真实温度 $T_\star$。此时，她仍然可以写出后验分布，但需要根据数据估计温度 T。一种做法是找到T 使分类误差最小的值，或者等价地，使泛化损失最小的值；不过，这也等价于采用期望最大化，相关讨论可参见 [Decelle 等人，2011年，Krzakala 等人， 2012]。这样的温度估计将带来 $T = T_*$ 并恢复贝叶斯最优估计器 f_{bo}。因此，在设定正确的贝叶斯设定中，进行温度缩放就相当于实现期望一致性，从而为贝叶斯框架下的温度缩放和期望一致性方法提供了十分自然的解释。

请注意，本文关注的是通过经验风险最小化训练的频率学派估计器。在这种情况下，即使在设定正确的情形下，TS 和 EC 在高维极限下也都无法恢复正确的温度 T_*。这种无法恢复 T_* 的情况源于：我们不再对某个分布进行采样，而是考虑一个点估计，并且样本数量不足以达到点估计器具有一致性的范围。

5.2 失配经验风险最小化

现在考虑这样一种情况：统计学家只能获取训练数据 $\mathcal{D}$，而不了解底层生成模型。在这种情况下，二元分类中一种常用的分类器是逻辑回归，其形式如下：

$$\hat{f}_{\mathrm{erm}}(\boldsymbol{x}) = \sigma(\hat{\boldsymbol{w}}^\top \boldsymbol{x}) \tag{8}$$

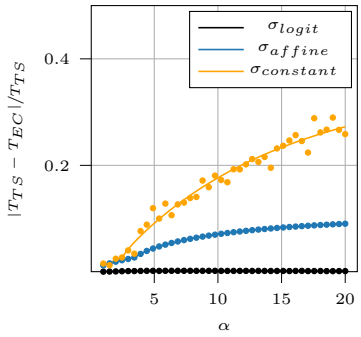


图5：相对差异 $|T_{EC} - T_{TS}|/T_{TS}$ 随采样比率 α 变化的情况，其中采用了三种不同的 $\sigma_\star$，以及 $\lambda = 10^{-4}$。我们观察到，当 $\sigma_\star$ 与 σ的差异增大时，期望一致性和温度缩放会得出不同结果。点表示在 $d = 200$进行的模拟。

ˆ，并且通过在训练数据 ˆ$\boldsymbol{w}$ = 上最小化经验风险 $\mathcal{R}_n$argmin$\boldsymbol{w}$ 来获得权重，其中：

$$\hat{\mathcal{R}}_n(\boldsymbol{w}) = -\sum_{i\in[n]} \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}) + \lambda/2\|\boldsymbol{w}\|^2 \tag{9}$$

并提醒一下， σ 是 Sigmoid/Logit函数。在此设定下，校准由 $\Delta_\ell = \ell - \mathbb{E}_{\boldsymbol{x}}\left[f_*(\boldsymbol{x})|\,\hat{}f_{\text{erm}}(\boldsymbol{x}) = \ell\right]$给出，期望校准误差则由 $\mathbb{E}_{\boldsymbol{x}}\left[|\Delta_{\hat{f}_{\text{erm}}}(\boldsymbol{x})|\right]$给出。

请注意，逻辑回归也可以看作Logit模型的最大似然估计量。根据数据模型（6）， $\sigma_\star \neq \sigma$ 的设定是错误的。Sur 和 Cand`es [2019] 指出，即使在设定正确情形下， $\sigma_\star = \sigma$未正则化的逻辑回归也会产生有偏的 $\boldsymbol{w}_\star$ 估计量，在高维极限下，当 n， $d \to \infty$ 以成比例的速率 $\alpha = n/d$增长时；Bai等人 [2021] 证明了该估计量具有过度自信的问题。Clart´e 等人 [2022b] 将校准刻画为正则化强度和样本数量的函数，并证明通过适当的正则化可以减轻过度自信。

本节旨在利用这些关于高维逻辑回归的结果，为温度缩放和期望一致性的校准性质提供理论结果。具体而言，我们将比较以下三种数据似然函数的选择 $\sigma_\star$：

$$\begin{cases} \sigma_{\text{logit}}(z) = \frac{1}{1+e^{-z}} \\ \sigma_{\text{affine}}(z) = 0 \text{ if } z < -1, 1 \text{ if } z > 1, \frac{t+1}{2} \text{ else} \\ \sigma_{\text{constant}} = 0 \text{ if } z < -1, 1 \text{ if } z > 1, 1/2 \text{ else} \end{cases} \tag{10}$$

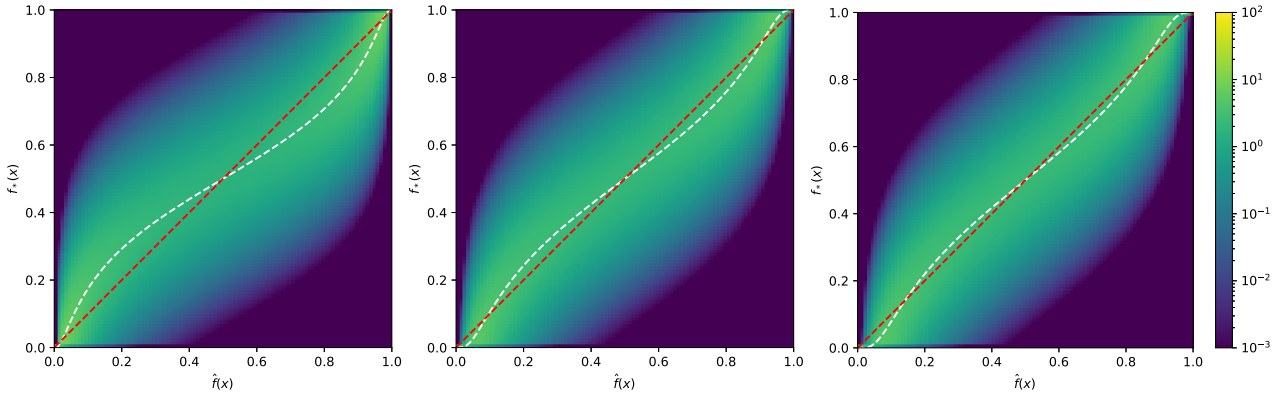
渐近不确定性度量 — 分析的出发点是注意到，我们感兴趣的不确定性度量（2）仅通过测试点 $(\boldsymbol{w}_\star^\top \boldsymbol{x},$ ˆ $\boldsymbol{w}^\top \boldsymbol{x})$ 上的预激活值 $\boldsymbol{x}$依赖于权重。由于输入服从高斯分布，预激活值的联合统计量也服从高斯分布：

$$(\boldsymbol{w}_\star^\top \boldsymbol{x}, \hat{\boldsymbol{w}}^\top \boldsymbol{x}) \sim \mathcal{N}\left(\boldsymbol{0}_2, \begin{bmatrix} 1/d||\boldsymbol{w}_\star||_2^2 & 1/d\boldsymbol{w}_\star^\top \hat{\boldsymbol{w}}_{\text{erm}} \\ 1/d\boldsymbol{w}_\star^\top \hat{\boldsymbol{w}}_{\text{erm}} & 1/d||\hat{\boldsymbol{w}}_{\text{erm}}||_2^2 \end{bmatrix}\right)$$

如上所述，近期的不同研究 [Sur 和 Cand`es，2019年，Bai 等人，2021年，Clart´e 等人，2022b] 针对逻辑回归，在不同的一般性层次下推导出了这些统计量的精确渐近公式。具体而言，分析将采用以下由 Clart´e 等人 [2022b], 提出的定理，该定理考虑了一个一般的模型设定错误情形：

定理 1 (来自 Clart´e 等人 [2022b]的定理3.2)。 考虑逻辑函数分类器 (8)， 该分类器通过最小化数据集 (9) 上的经验风险进行训练；该数据集 $(\boldsymbol{x}_i, y_i)^{i\in[n]}$ 独立采样自模型 (6)。那么，在高维极限下， 当 $n, d \to \infty$ 取固定值 $\alpha = n/d$时：

$$(1/d\boldsymbol{w}_\star^\top \hat{\boldsymbol{w}}_{\text{erm}}, 1/d||\hat{\boldsymbol{w}}_{\text{erm}}||_2^2) \xrightarrow[d\to\infty]{} (m, q) \tag{11}$$



ˆ 图 6：对于采样比率 $n/d = 20$ 和正则化 $\lambda = 10^{-4}$，$(f_{\text{erm}}(\boldsymbol{x})\ f_\star(\boldsymbol{x}))$ 的密度图（左）、经过温度缩放后的密度图（中）以及经过期望一致性处理后的密度图（右）。白色虚线表示准确率随置信度变化的函数，红线为对角线。红线ˆ与白线之间的差异对应于校准。ECE 从左到右分别为：2.1%、1.2%、1.0%。我们有$T_{TS} = 1.24$，$T_{EC} = 1.35$。

其中，$(m, q) \in \mathbb{R}^2_+$ 由一组低维自洽方程的解显式给出，这组方程仅取决于$(\alpha, \lambda, \sigma, \sigma_*)$；为节省篇幅，相关讨论见附录 B。

借助定理 1，我们可以推导出式（2）中定义的不确定性指标的渐近极限的渐近刻画。

命题 1。 在与定理 *1*相同的假设下，式 （2）中定义的不确定性指标的渐近极限为：

$$\begin{cases} \Delta_\ell(m,q) & = \ell - \mathcal{Z}_\star(1, {}^m\!/_q \sigma^{-1}(\ell), 1 - {}^{m^2}\!/_q) \\ \text{ECE}(m,q) & = \int_0^\infty \mathrm{d}z |\Delta_{\sigma(z)}(m,q)| \mathcal{N}(z|0,q) \end{cases}$$

其中 $(m、\ q) \in \mathbb{R}^2_+$ 是 (11) 中相关函数的渐近极限，并且

$$\mathcal{Z}_\star(y, \omega, V) = \mathbb{E}_{\xi \sim \mathcal{N}(\omega, V)} \left[\sigma_\star \left({}^{y\xi}\!/_{T_\star}\right)\right] \tag{12}$$

该结果的证明见附录 B。命题 1 为我们完整刻画当前设定下 TS 和 EC 的校准性质提供了所需的一切。接下来几段将讨论其影响。

在实践中，经验风险（ ℓ_2 正则化参数 λ （9）通常通过交叉验证进行优化。Clart´e 等人 [2022b、a]表明，适当地对风险进行正则化不仅能提高预测准确率，还能改善逻辑分类器的校准和 ECE。具体而言，研究表明，在损失函数上进行交叉验证所得的结果不同于在误分类误差上进行交叉验证所得的结果；对于模型设定错误的情形，两者的差异更大。值得注意的是，Clart´e 等人 [2022a] 表明，在这种情况下，将一个 ℓ_2 惩罚项与 TS 结合，能够实现良好的性能和校准效果。下文将讨论这与 EC 的比较。注意，定理 1 给出的精确渐近刻画使我们能够绕过交叉验证，通过直接优化低维公式来找到最优的 λ。因此，我们定义 λ_{error} （分别为 λ_{loss}）为 λ 使得 $\boldsymbol{w}_{\text{erm}}$ 取得最低测试误分类误差（分别为测试损失）的值。

5.3 期望一致性优于温度缩放

在第 4 节中，我们通过数值实验观察到，对于在真实数据集上训练的不同架构，EC 和 TS 得出的温度几乎相同，因此在不同不确定性量化指标上的表现也相近。图 5 展示了在合成数据模型（ $\delta T = {}^{|T_{TS} - T_{EC}|}\!/_{T_{TS}}$ ）上，逻辑回归采用不同目标激活函数时这两种方法的相对差异（6）

$\sigma_\star \in \{\sigma_{\text{logit}}, \sigma_{\text{affine}}, \sigma_{\text{constant}}\}$ 定义见（10）。与第 4 节中的真实数据场景相反，我们观察到这两种方法之间存在显著差异，适用于 $\sigma_\star \in \{\sigma_{\text{affine}}, \sigma_{\text{constant}}\}$。例如，对于分段常数函数 $\sigma_\star = \sigma_{\text{constant}}$， δT 是采样比率 α的非递减函数，在 30%处为 $\alpha = 20$。

图4 显示，在第5节所考察的所有设定中，期望一致性产生的期望校准误差都低于温度缩放。一方面，在目标似然与模型似然相同的设定正确情形下，这种效果较小：温度缩放的期望校准误差高出约 0。01%。根据第5.1节的讨论，这相当直观，因为在这种情况下，我们更接近贝叶斯设定，而在该设定下，已证明这两种方法等价。另一方面，在设定错误情形下，这一差异会增大，说明模型设定错误对这些校准方法有重要影响。特别要注意的是，在此处考察的所有情形中，对于所考虑的三种正则化项，期望一致性的期望校准误差都低于温度缩放：$\lambda = 10^{-4}$， λ_{error}， λ_{loss}。

图6 显示了变量的联合概率密度函数 $f_*(\boldsymbol{x})$，$\hat{f}_{\text{erm}}(\boldsymbol{x})$ 。$[0,1]^2$ 上概率密度函数用白色虚线表示条件均值 $\mathbb{E}\left[f_*(\boldsymbol{x}) | \hat{f}_{\text{erm}}(\boldsymbol{x})\right]$，它对应于图2 中的准确率-置信度图。与真实数据情形一样，我们观察到经验风险最小化估计量始终过度自信，即 $\forall \ell \geqslant 1/2$， $\Delta_\ell \geqslant 0$。此外，我们看到，经过温度缩放和期望一致性处理后，条件均值更接近对角线（红色曲线），这表明模型的校准效果更好。简单数据模型呈现的现象似乎与真实数据中的观察结果相符，并表明期望一致性是更好的校准方法。

结果解读——温度缩放通过最小化验证损失，对网络的输出进行重新缩放。在相关文献中，交叉熵损失是最广泛使用的选择之一，既用于训练，也用于衡量不确定性分数（结合 Softmax）。从贝叶斯的角度来看，最小化交叉熵损失等价于在假设数据由 Softmax（又称多项逻辑）模型生成的情况下最大化似然。因此，温度缩放背后的基本假设是：标签由 Softmax 似然生成。由此，我们预计，当这一假设成立时，温度缩放的表现会更好。事实上，我们在第 5 节中的实验验证了这一直觉。在真实模型确实由逻辑函数给出的情况下，TS 表现良好，其结果接近期望一致性（EC）。然而，在这一假设不成立的模型设定错误情况下，TS 的表现不如 EC。

6 结论与未来工作

在这项工作中，我们介绍了期望一致性，一种新的神经网络后训练校准方法。我们证明，在不同的图像分类任务中，EC 与温度缩放的表现十分接近：二者具有几乎相同的期望校准误差和布里尔分数，同时计算成本也相当。此外，我们还分析了这两种方法在合成环境下的渐近性质。在该环境中，数据由真实模型生成。分析表明，对于设定正确的方法，EC 和 TS 能够取得相同的性能；而在模型设定错误的情况下，EC 提供了更优且更具原则性的校准方法。

我们在简单数据模型上的实验表明，当线性模型与真实数据模型之间存在差异时，EC 的表现优于 TS。然而，我们在真实数据上的实验显示，在不同架构、数据模型和整体模型准确率下，各方法的表现非常相似。未来，我们希望进一步了解为何这两种方法在实际场景中的表现如此相近。

7 致谢

我们感谢欧盟“地平线2020”研究与创新计划资助协议714608-SMiLe项下的欧洲研究理事会（ERC）资助、瑞士国家科学基金会资助（SNFS OperaGOST、200021 200390），以及选择法国——法国国家科学研究中心（*CNRS*）-*AI* 新锐人才计划的资助。本研究还得到了

NCCR MARVEL（一个国家研究能力中心），由瑞士国家科学基金会资助（资助编号：205602）。

参考文献

Moloud Abdar、Farhad Pourpanah、Sadiq Hussain、Dana Rezazadegan、Li Liu、Mohammad Ghavamzadeh、Paul Fieguth、Xiaochun Cao、Abbas Khosravi、U. Rajendra Acharya、Vladimir Makarenkov 和 SaeidNahavandi。《深度学习中不确定性量化综述：技术、应用与挑战》。信息融合，76:243–297，2021年。

Yu Bai、Song Mei、Huan Wang 和 Caiming Xiong。别只把过度自信归咎于过参数化：二元分类中校准的理论分析。载于 Marina Meila 和 Tong Zhang（编），第*38*届国际机器学习会议论文集，第139卷，机器学习研究论文集，第566–576页。PMLR，2021年7月18–24日。

Lucas Clart´e、Bruno Loureiro、Florent Krzakala 和 Lenka Zdeborov´a。《过参数化高维模型中的不确定性量化研究》，2022a。

Lucas Clart´e、Bruno Loureiro、Florent Krzakala 和 Lenka Zdeborov´a。高维线性分类中不确定性的理论刻画，2022b。

Aurelien Decelle、Florent Krzakala、Cristopher Moore 和 Lenka Zdeborov´a。模块化网络随机块模型的渐近分析及其算法应用。《物理评论*E*》，84(6)：066106，2011年。

Arthur P Dempster、Nan MLaird 和 Donald BRubin。通过EM算法从不完整数据中获得最大似然。《皇家统计学会杂志：*B*辑（方法论）》，39(1)：1–22，1977年。

Xiaohan Ding、Xiangyu Zhang、Ningning Ma、Jungong Han、Guiguang Ding 和 Jian Sun。RepVGG：让 VGG 风格卷积网络再次强大。在*IEEE/CVF*计算机视觉与模式识别会议论文集中，第13733–13742页，2021年。

Lior Frenkel 和 Jacob Goldberger。基于类别的温度缩放实现网络校准。载于*2021*年 第*29*届欧洲信号处理会议（*EUSIPCO*），第1486–1490页，2021年。

Yarin Gal 和 Zoubin Ghahramani。将随机失活作为贝叶斯近似：表示深度学习中的模型不确定性。载于第*33*届国际机器学习会议论文集——第*48*卷，2016年。

Jakob Gawlikowski、Cedrique Rovile Njieutcheu Tassi、Mohsin Ali、Jongseok Lee、Matthias Humt、Jianxiang Feng、Anna Kruspe、Rudolph Triebel、Peter Jung、Ribana Roscher、Muhammad S hahzad、Wen Yang、Richard Bamler 和 Xiao Xiang Zhu。深度神经网络不确定性综述，2021年。

Alex Graves。神经网络的实用变分推断。载于 J. Shawe-Taylor、R. Zemel、P. Bartlett、F. Pereira 和 K.Q. Weinberger 编，神经信息处理系统进展，第24卷。Curran Associates, Inc.，2011年。

Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q. Weinberger。现代神经网络的校准。载于 Doina Precup 和 Yee Whye Teh 编，《第34届国际机器学习会议论文集》，第70卷《机器学习研究论文集》，第1321–1330页。PMLR，2017年8月6日至11日。

Chirag Gupta、Aleksandr Podkopaev 和 Aaditya Ramdas。无分布假设二分类：预测集、置信区间和校准。载于《第34届神经信息处理系统国际会议论文集》，NIPS'20，美国纽约州雷德胡克，2020年。Curran Associates Inc.

K. He、X. Zhang、S. Ren 和 J. Sun。用于图像识别的深度残差学习。载于*2016*年*IEEE*计算机视觉与模式识别会议（*CVPR*），第 770–778 页，洛斯阿拉米托斯，加利福尼亚州，美国，2016年6月。IEEE计算机学会。

G. Huang、Z. Liu、L. Van Der Maaten 和 K. Q. Weinberger。密集连接卷积网络。载于*2017*年*IEEE*计算机视觉与模式识别会议（*CVPR*），第 2261–2269 页，洛斯阿拉米托斯，加利福尼亚州，美国，2017年7月。IEEE计算机学会。

Yukito Iba。西森线与贝叶斯统计。*Journal ofPhysics A: Mathematical andGeneral*，32(21): 3875，1999年。

Agustinus Kristiadi、Matthias Hein 和 Philipp Hennig。只要稍微具备贝叶斯特性，就能消除ReLU网络的过度自信。载于 Hal Daum´e III 和 Aarti Singh 编，第*37*届国际机器学习会议，第119卷，机器学习研究论文集，第5436–5446页。PMLR，2020年7月13–18日。

Alex Krizhevsky。学习小图像中的多层特征。载于技术报告，多伦多大学，2009年。

Florent Krzakala、Marc M´ezard、Francois Sausset、Yifan Sun 和 Lenka Zdeborov´a。压缩感知中的概率重构：算法、相图与达阈值矩阵。*Journal ofStatistical Mechanics: Theory andExperiment*，2012(08):P08009，2012年。

Meelis Kull、Miquel Perello-Nieto、Markus K¨angsepp、Telmo Silva Filho、Hao Song 和 Peter Flach。《超越温度缩放：使用狄利克雷校准获得良好校准的多类别概率》。载于《第33届神经信息处理系统国际会议论文集》，美国纽约州雷德胡克，2019年。Curran Associates Inc.

Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。《使用深度集成进行简单且可扩展的预测不确定性估计》。载于《第31届神经信息处理系统国际会议论文集》，NIPS'17，第6405–6416页，美国纽约州雷德胡克，2017年。Curran Associates Inc.

Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。《使用深度集成进行简单且可扩展的预测不确定性估计》。载于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vish-wanathan 和 R. Garnett 编，《神经信息处理系统进展》，第30卷。Curran Associates, Inc.，2017年。

Bruno Loureiro、Cedric Gerbelot、Hugo Cui、Sebastian Goldt、Florent Krzakala、Marc Mezard 和 LenkaZde- borov´a。《真实数据集上采用师生模型的通用特征映射学习曲线》。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，《神经信息处理系统进展》，第 34 卷，第 18137–18151 页。Curran Associates, Inc.，2021年。

Wesley J. Maddox、T. Garipov、Pavel Izmailov、Dmitry P. Vetrov 和 Andrew Gordon Wilson。《深度学习中贝叶斯不确定性的简单基线方法》。载于*NeurIPS*，2019年。

Pierre-Alexandre Mattei。《贝叶斯模型不确定性的简约概览》，2019年。

Cyril M´easson、Andrea Montanari、Thomas J Richardson 和 R¨udigerUrbanke。《广义面积定理及其若干推论》。*IEEE*信息论汇刊，55(11)：4793–4821，2009年。

Matthias Minderer、Josip Djolonga、Rob Romijnders、Frances Hubis、Xiaohua Zhai、Neil Houlsby、Dustin Tran 和 Mario Lucic。重新审视现代神经网络的校准。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P. S. Liang 和 J. Wortman Vaughan（编辑），神经信息处理系统进展，第34卷，第15682–15694页。Curran Associates, Inc.，2021年。

Yuval Netzer、Tao Wang、Adam Coates、Alessandro Bissacco、Bo Wu 和 Andrew Y. Ng。使用无监督特征学习读取自然图像中的数字。载于 *NIPS*深度学习与无监督特征学习研讨会 *2011*，2011年。

Harris Papadopoulos、Kostas Proedrou、Volodya Vovk 和 Alex Gammerman。用于回归的归纳置信机。载于 机器学习：*ECML 2002*：第*13*届欧洲机器学习会议，芬兰赫尔辛基，*2002*年*8*月*19*日至*23*日，会议论文集*13*，第345–356页。施普林格，2002年。John Platt。支持向量机的概率输出及其与正则化似然方法的比较。大间隔分类进展，10，2000年6月。Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*CoRR*，abs/1409.1556，2014年。Pragya Sur 和 Emmanuel J. Cand`es。高维逻辑回归的现代最大似然理论。美国国家科学院院刊，116(29):14516–14525，2019年。Deng-Bao Wang、Lei Feng 和 Min-Ling Zhang。重新思考深度神经网络的校准：不要害怕过度自信。载于神经信息处理系统，2021年。Bianca Zadrozny 和 Charles Elkan。从决策树和朴素贝叶斯分类器中获得校准概率估计。国际机器学习大会，1，2001年5月。Lenka Zdeborov´a 和 Florent Krzakala。推断的统计物理学：阈值与算法。物理学进展，65(5):453–552，2016年。

A 训练过程详情

SVHN 对于 SVHN 数据集，Netzer 等人 [2011], 训练了深度为 20 且包含 0.27M 个参数的 ResNet-20 模型，共进行 50 个周期，使用随机梯度下降，学习率为 $\eta = 0.1$，权重衰减为 $1e-4$，动量为0。9。 90% 的数据点用于训练，其余用于验证。

CIFAR10 对深度为 20 和 56 的残差网络模型（其中 ResNet-56 包含 0.85M 个参数）进行了训练，共 50个周期，使用随机梯度下降，学习率为 $\eta = 0.1$，权重衰减为 $1e-4$，动量为 0。9。包含 7.9 个参数的 DenseNet-121 使用与残差网络模型相同的参数进行训练，但学习率除外$\eta = 0.01$。如 He 等人 [2016], 所述，训练集中的图像经过随机裁剪并进行水平翻转。

CIFAR100 在 CIFAR100 上，我们使用了来自 GitHub 仓库 https://github.com/chenyaofo/pytorch-cifar-models的预训练模型。这些模型在完整的训练集上进行训练，因此包含 10000 张图像的测试集被均分为验证集和测试集，分别包含 5000 张图像。

A.1 补充图表

在图 7中，我们分别绘制了 SVHN 和 CIFAR10 上 ResNet-20 与 ResNet-56 的可靠性图。我们观察到，未校准的模型过度自信（因为置信度高于相应的准确率），而 TS 和 EC 均能缓解这种过度自信。

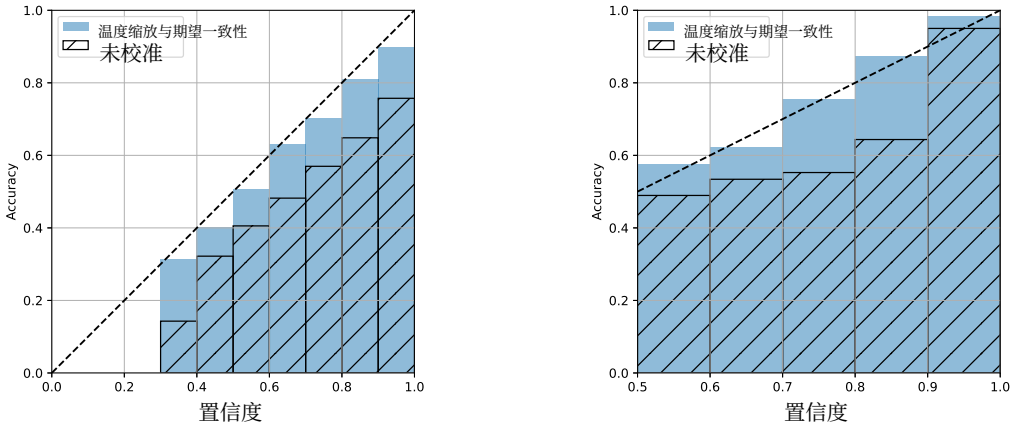


图 7：SVHN 数据集上 ResNet-20（左）和 CIFAR10 数据集上 ResNet-56（右）的可靠性图。在校准之前，两种方法都过度自信。TS 和 EC 改善了校准效果，并缓解了过度自信。

B 状态演化方程

本节重点讨论第5节中引入的数据模型。回顾一下，我们考虑的是一个包含 n个样本 $\mathcal{D} = (x^\mu, y^\mu)_{\mu=1}^n$ 的数据集，该数据集由

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_d/d), \boldsymbol{w}_* \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_d), \mathbb{P}(y=1|\boldsymbol{w}_*^\top \boldsymbol{x}) = \sigma_\star(\boldsymbol{w}_*^\top \boldsymbol{x}) \tag{13}$$

并拟合以下逻辑回归模型，其中 σ 包含 Sigmoid函数:

$$\hat{f}_{\text{erm}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x}) \tag{14}$$

通过最小化以下经验风险

$$\mathcal{R}(\boldsymbol{w}, \mathcal{D}, \lambda) = \sum_{\mu=1}^{n} \log \sigma(y^\mu \boldsymbol{w}^\top \boldsymbol{x}) + \lambda/2 \|\boldsymbol{w}\|^2 \tag{15}$$

因此我们有 $\boldsymbol{w}_{\text{erm}} = \arg\min_{\boldsymbol{w}} \mathcal{R}(\boldsymbol{w},\ \mathcal{D},\ \lambda)$。对于一个新样本 $\boldsymbol{x}$，我们关注 $f_*(\boldsymbol{x})$ 和 $\hat{} f_{\text{erm}}(\boldsymbol{x})$ 的联合分布。由于这两个函数仅取决于标量积 $\boldsymbol{w}_*^\top \boldsymbol{x}, \boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x}$，因此只需计算这些标量的联合分布。根据 $\boldsymbol{x}$ 的高斯性，我们只需计算重叠 $m = \boldsymbol{w}_*^\top \boldsymbol{w}_{\text{erm}}$ 和 $q = \|\boldsymbol{w}_{\text{erm}}\|^2$。在 $n,\ d \to \infty$ 的高维极限下，同时保持采样比率 $n/d = \alpha$ 不变，可以计算出 m 和 q 的值。其思路是引入分布

$$\mu_{\beta,\mathcal{D},\lambda}(\boldsymbol{w}) = \frac{1}{\mathcal{Z}_\beta} \exp\left(-\beta \mathcal{R}(\boldsymbol{w}, \mathcal{D}, \lambda)\right) \tag{16}$$

其中 $\mathcal{Z}_\beta$ 是一个归一化常数。在极限 $\beta \to \infty$ 下，$\mu_{\beta\mathcal{D}\lambda}$ 收敛到一个以 $\boldsymbol{w}_{\text{erm}} = \arg\min \mathcal{R}(\boldsymbol{w}, \mathcal{D}, \lambda)$ 为峰值的狄拉克分布。为了计算 mq，需要求出 $\log\ \mathcal{Z}_\beta$ 的表达式及其在 $\beta \to \infty$ 时的极限。在维度和样本数量均趋于无穷且比值固定的高维情形下，可以使用统计物理学中的副本方法进行计算，参见 Zdeborov´a 和 Krzakala[2016]。由于这些计算并非本文的重点，详细计算过程请参见 Loureiro 等人 [2021], Clart´e 等人 [2022b] 。最后，若定义

$$\mathcal{Z}_*(y, \omega, v_*) = \int \mathrm{d}z \sigma_*(y \times z) \mathcal{N}(z|\omega, v_*) \tag{17}$$

$$f(y, \omega, v) = \arg\min_z \left[\frac{(z-\omega)^2}{2v} - \log \sigma(z)\right] \tag{18}$$

则 $m,\ q$ 是以下自洽方程的解：

$$\begin{cases} m &= \frac{\hat{m}}{\lambda+\hat{v}} \\ q &= \frac{\hat{q}+\hat{m}^2}{(\lambda+\hat{v})^2} \\ v &= \frac{1}{\lambda+\hat{v}} \end{cases}, \quad \begin{cases} \hat{m} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\int \mathrm{d}y \partial_\omega \mathcal{Z}_*(y, m/q\xi, v_*) f(y, \xi, v)\right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\int \mathrm{d}y \mathcal{Z}_*(y, m/q\xi, v_*) f^2(y, \xi, v)\right] \\ \hat{v} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\int \mathrm{d}y \mathcal{Z}_*(y, m/q\xi, v_*) \partial_\omega f(y, \xi, v)\right] \end{cases} \tag{19}$$

其中 $v_* = \rho - m^2/q$。

高维情形下的校准 得到重叠量 $m,\ q$后，我们可以推导出校准 Δ_ℓ的表达式：

$$\Delta_\ell = \mathbb{E}\left[f_*(\boldsymbol{x})|\hat{f}_{\text{erm}}(\boldsymbol{x})\right] = \mathbb{P}\left[y = 1|\hat{f}_{\text{erm}}(\boldsymbol{x})\right] = \int \mathrm{d}z \sigma_\star(z) \mathcal{N}(z|\frac{m}{q}\hat{f}_{\text{erm}}^{-1}(\boldsymbol{x}), \rho - m^2/q) \tag{20}$$

第二行源于以下事实：在给定 $\boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x} = \sigma^{-1}(\ell)$ 的条件下，标量积 $\boldsymbol{w}_*^\top \boldsymbol{x}$ 服从均值为 $m/q\xi$、方差为 $\rho - m^2/q$ 的高斯分布。因此，ECE 的表达式为

$$ECE = \mathbb{E}_{\boldsymbol{x}}\left[|\Delta_{\hat{f}_{\text{erm}}(\boldsymbol{x})}|\right] = \mathbb{E}_{\xi=\boldsymbol{w}_{\text{erm}}^\top \boldsymbol{x}}\left[|\Delta_{\sigma(\xi)}|\right] = \int \mathrm{d}\xi |\Delta_{\sigma(\xi)}| \mathcal{N}(\xi|0, q) \tag{21}$$

C 腐化数据集上的实验

下面介绍一项实验，在该实验中，EC 在真实数据上的表现可显著优于 TS：我们按照图 1 所示，在若干图像分类任务上训练不同的架构。不过，这里验证集和测试集中有些类别被替换为随机标签。对于 SVHN 和 CIFAR10，标签 $y = 0$ 被替换为随机标签。对于 CIFAR100，标签 $y = 0$、 ..... 、 9 被替换为随机标签。这样，约 10% 的验证/测试数据被污染，且噪声取决于类别。请注意，训练数据保持不变：本实验的目的是模拟训练数据与测试数据之间的分布偏移，类似于? 中所做的那样。

在下表中，我们比较了 EC 和 TS 在这些受污染的数据集上的表现（以期望校准误差和 Brier评分（BS）衡量）。我们观察到，在这种设置下，EC 在多个数据集和架构上的表现显著优于 TS。

数据集	模型	$\mathcal{E}_g$	T_{TS}	T_{EC}	ECE	ECE_{TS}	ECE_{EC}	BS	BS_{TS}	BS_{EC}
SVHN	ResNet-20	12.5 %	2.69	2.23	8.3 %	10.7 %	7.5 %	21.9 %	23.4 %	22.1 %
CIFAR10	ResNet-20	20.9 %	2.4	2.0	12.8 %	4.6 %	4.2 %	34.2 %	32.2 %	32.1 %
CIFAR10	ResNet-56	21 %	2.58	2.15	13.8 %	5.4 %	4.9 %	35.2 %	32.9 %	32.8 %
CIFAR10	DenseNet-121	20.4 %	2.76	2.54	15.8 %	3.6 %	5.0 %	35.9 %	31.8 %	31.9 %
CIFAR100	ResNet-20	38.1 %	2.04	1.70	16.5 %	9.6 %	5.9 %	57.0 %	54.9 %	53.9 %
CIFAR100	ResNet-56	34.8 %	2.27	2.10	21.7 %	7.6 %	7.3 %	56.0 %	50.6 %	50.4 %
CIFAR100	VGG19	35.5 %	2.6	2.1	28.34 %	5.2	5.1 %	61.8 %	50.1 %	50.1 %
CIFAR100	RepVGG-A2	30.5 %	1.44	1.40	13.7 %	11.6 %	11.7 %	47.2 %	47.1 %	47.0 %

表1：当部分验证数据和测试数据遭到破坏时，温度缩放（TS）与期望一致性（EC）的期望校准误差（ECE）和Brier 分数（BS）比较