{Re}

高效回归模型的符合预测方法在标签噪声下的应用

Yahav Cohen, Jacob Goldberger and Tom Tirer

Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel

ABSTRACT

在高风险场景中，如医学成像应用，为回归模型的预测配备可靠的置信区间至关重要。最近，一致预测（CP）作为一种强大的统计框架出现，基于标记校准集生成包含真实标签的预设概率的区间。本文解决了当校准集中包含噪声标签时将 CP 应用于回归模型的问题。我们首先建立了一个以数学为基础的过程来估计无噪声的 CP 阈值。然后，我们将这一过程转化为一个实际算法，克服了由于回归问题连续性所带来的挑战。我们在两个具有高斯标签噪声的医学成像回归数据集上评估了所提出的方法。我们的方法显著优于现有替代方案，达到了接近干净标签设置的表现。

Index Terms— 回归神经网络，符合性预测，标签噪声

1. 介绍

回归深度神经网络（DNN）在现代机器学习中扮演着关键角色，作为从复杂高维数据预测连续值的系统的核心。它们广泛应用于各个领域：从医学图像估计患者的解剖指标，预测能源消耗和金融趋势，通过轨迹预测引导自动驾驶汽车等 [1–4]。

对于安全关键的应用，仅准确预测是不够的，报告预测的信心对于可靠和可解释的模型至关重要。一种广泛采用的方法是通过置信区间传达回归 DNN 预测的不确定性，这些区间应以预定义的概率包含真实值。这些区间的大小预计会很小，并与案例的复杂性相关联。

最近，一致性预测（CP）[5–7] 作为实现此目的的强大通用统计框架而崭露头角。基于带有标签的校准集，CP 在数据分布（独立同分布或更一般地说，校准和测试样本的可交换性）非常温和的假设下返回具有覆盖保证的信心区间（分类情况下为“预测集”）。具体来说，信心区间的真正标签包含概率是预先指定的。目标是在保持覆盖水平的同时返回最小的区间，并根据不同 CP 方法的信心区间或“预测集” [7–9] 的平均长度来判断它们，在文献中也称为效率。CP 已成为医疗成像 [10–12] 等安全关键应用中的一个重要校准工具。最新的关于回归应用的一致性预测综述是 [13, 14]。

CP 在诸如医学成像等应用中面临的关键挑战来自标签噪声。在这些领域，数据集中经常包含源自模糊数据的嘈杂标签，这可能会混淆甚至临床专家。此外，医生可能对同一医学图像的诊断意见不一，导致地面真实标签存在不一致。基于嘈杂标签校准模型的挑战直到最近才开始受到关注，主要集中在分类设置上。Einbinder 等人 [15] 建议忽略标签噪声并简单地将标准 CP 算法应用于嘈杂标记的校准集。对于回归而言，这种策略会导致更大的置信区间。其他相关工作 [16–19] 提供了对 CP 针对嘈杂标签的适应，但重点是分类。其中一些展示了覆盖保证界限。然而，在许多情况下这些界限过于保守（导致预测集合非常大）。

在这项研究中，我们首次尝试利用带噪声标签的校准数据集将 CP 应用于回归 DNN。我们提出了一种新的算法来估计无噪声的 CP 阈值，该算法通过建立一个数学上成立的过程并将其转化为实用方法以克服由回归问题连续性带来的挑战而开发。我们还讨论了在模型训练期间没有干净标签的情况。我们在带有高斯标签噪声的两个医学成像回归数据集上评估了所提出的方法。我们的方法对噪声水平具有鲁棒性，有效覆盖，并且产生的平均区间长度（效率）接近于无噪声标签设置，明显短于 [15] 方法得到的结果，这促使忽略标签噪声。

2. 背景

令 (X, Y) 表示一个样本及其标签，分布于 $\mathcal{X} \times \mathcal{Y}$ 上。对于具有标量标签的回归任务，我们有 $\mathcal{Y} = \mathbb{R}$。考虑一个 DNN 对输入样本 $x \in \mathcal{X}$ 输出预测值 $\hat{y}(x) \in \mathcal{Y}$，校准集上的带标签样本 $\{x_i, y_i\}_{i=1}^n$，以及预定义的 $\alpha \in (0, 1)$。

一致性预测（CP）建立了一种决策规则，用于生成置信区间 $x \mapsto \mathcal{C}(x)$，使得 $Y \in \mathcal{C}(X)$ 的概率为 $1-\alpha$，其中 Y 是与 X [5,6] 相关的真实类别。一致性预测中的唯一假设是校准集和测试样本相关的随机变量是可交换的（例如，样本是独立同分布的）。让我们陈述一般的 CP 框架 [7]：

1. 定义一个基于模型某些输出的启发式得分函数 $s(x, y) \in \mathbb{R}$。更高的分数应表示 x 和 y 之间较低的一致性水平。

2. 校准阶段：计算 $\hat{q}$ 作为分数 $\{s(x_1, y_1), \ldots, s(x_n, y_n)\}$ 的 $\lceil (n+1)(1-\alpha) \rceil / n$ 分位数。

3. 部署阶段：使用 $\hat{q}$ 创建新样本 x_{n+1} 的预测集：$\mathcal{C}(x_{n+1}) = \{y : s(x_{n+1}, y) \leq \hat{q}\}$。

在回归情况下，其中 $\mathcal{Y} = \mathbb{R}$，本质上我们得到了置信区间 $C(x_{n+1}) \subset \mathbb{R}$。CP 方法具有以下覆盖保证。

Theorem 1 (**定理 1 在 [7] 中**) 假设 $\{(X_i, Y_i)\}_{i=1}^n$ 和 (X_{n+1}, Y_{n+1}) 是独立同分布的。定义 $\hat{q}$ 如下述步骤 2 中的定义，$\mathcal{C}(X_{n+1})$ 如下述步骤 3 中的定义。那么以下成立：$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$。

该结果的证明基于 [5]。$1 - \alpha + 1/(n+1)$ 的上界证明也存在。

不同的 CP 方法通常通过其选择的得分函数 $s(x, y)$ 有所不同，它们评判的一个关键属性是平均值 $|\mathcal{C}(x)|$（长度或 $\mathcal{C}(x)$ 的基数），通常称为效率。专注于回归，最简单的选择是：$s(x, y)$ 是 $s(x, y) = |y - \hat{y}(x)|$. 然而，由于：$\mathcal{C}(x) = \{y : |y - \hat{y}(x)| \leq \hat{q}\} \implies |\mathcal{C}(x)| = 2\hat{q}$，这会产生一个对任何样本都固定的区间，从而忽略了 x 是否是一个容易或困难的样本。请注意，回归 DNN 的输出 $\hat{y}(\cdot)$ 可以理解为后验均值的估计。通常会训练 DNN 也输出 $\hat{u}(\cdot)$，即后验标准差的估计，例如使用高斯负对数似然（NLL）损失（在实践中，输出 $\log(\hat{u}(\cdot))$ 有助于优化）。在这种情况下，一个能够提供更好效率的得分函数由 $s(x, y) = |y - \hat{y}(x)|/\hat{u}(x)$，给出，其置信区间为 $C(x) = \left[\hat{y}(x) - \hat{q}\hat{u}(x), \hat{y}(x) + \hat{q}\hat{u}(x)\right]$ [20]。为了简洁起见，本文也将使用这个得分及其导致的 $C(x)$。

在本文中，我们考虑校准集为 $\{x_i, \tilde{y}_i\}_{i=1}^n$ 的问题，其中 $\tilde{y}_i$ 是 y_i 的噪声版本。这种情况已在 [15] 中进行研究。作者已经证明，对于分散噪声而言，当分布 $\tilde{Y}|X$ 比分布 $Y|X$ 更为扩散时，在 $\{x_i, \tilde{y}_i\}_{i=1}^n$ 上执行标准的 CP 校准阶段将导致阈值 $\hat{q}$ 大于无噪声校准时获得的阈值，因此覆盖率得以保持：$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$，其中 Y_{n+1} 是测试样本 X_{n+1} 的干净标签。

不幸的是，正如之前在分类 [16–19] 中的工作所示，并将在回归中进一步展示，这种简单的被称为“噪声 CP”的方法会导致显著的覆盖率过高（超过所需的 $1-\alpha$），因此也会导致非常大的置信区间（效率低下）。

3. 带有标签噪声的回归中的共形预测

本文的目标是设计一种满足预定义覆盖率要求的同时，提供比“噪声 CP”显著更短置信区间的 CP 方法。为此，我们的方法将基于估计无噪声情况下的 CP 阈值。

3.1. 估计无噪声 CP 阈值

设 X，Y 和 $\tilde{Y}$ 分别为样本、其干净标签和其噪声标签的连续随机变量。设 $C_q(x)$ 是一个依赖于阈值 q 的置信区间，例如 $C_q(x) = [\hat{y}(x) - q\hat{u}(x), \hat{y}(x) + q\hat{u}(x)]$，与分数 $s(x, y) = |y - \hat{y}(x)|/\hat{u}(x)$ 相关联，对于一个输出预测 $\hat{y}(\cdot)$ 和启发式不确定性的模型 $\hat{u}(\cdot)$。我们做如下假设。

假设 A.1 给定 Y,$\tilde{Y}$ 和 X 是独立的。

假设 A.2 $p_{\tilde{Y}|Y}(\tilde{y}|y) = k(\tilde{y} - y; \sigma^2)$，即噪声可以表示为带宽参数为 σ^2 的核 k。

假设 A.1 在标签噪声 [16–19] 的文献中很常见。假设 A.2 包括例如加性高斯噪声的情况：$k(\cdot; \sigma^2) = \mathcal{N}(\cdot; 0, \sigma^2)$。

定义

$$M_q^c(\ell, y) := \mathbb{P}(\ell \in C_q(X)|Y = y)p_Y(y),$$
$$M_q^n(\ell, \tilde{y}) := \mathbb{P}(\ell \in C_q(X)|\tilde{Y} = \tilde{y})p_{\tilde{Y}}(\tilde{y}).$$

观察到

$$\begin{aligned}\mathbb{P}(Y \in C_q(X)) &= \int \mathbb{P}(\ell \in C_q(X)|Y = \ell)p_Y(\ell)\mathrm{d}\ell \\ &= \int M_q^c(\ell, \ell)\mathrm{d}\ell. \end{aligned} \tag{1}$$

也就是说，真实标签的覆盖概率取决于 M_q^c 的“迹”。然而，从带有噪声标记的校准集 $\{x_i, \tilde{y}_i\}_{i=1}^n$ 中，我们只能近似 M_q^n。这促使我们建立 M_q^n 和 M_q^c 之间的关系。确实，我们有以下结果：

$$\begin{aligned}&M_q^n(\ell, \tilde{y}) = \mathbb{P}(\ell \in C_q(X)|\tilde{Y} = \tilde{y})p_{\tilde{Y}}(\tilde{y}) \\ &= \int \mathbb{P}(\ell \in C_q(X)|\tilde{Y} = \tilde{y}, Y = y)p_{Y|\tilde{Y}}(y|\tilde{y})p_{\tilde{Y}}(\tilde{y})\mathrm{d}y \\ &= \int \mathbb{P}(\ell \in C_q(X)|\tilde{Y} = \tilde{y}, Y = y)p_Y(y)p_{\tilde{Y}|Y}(\tilde{y}|y)\mathrm{d}y \\ &= \int \mathbb{P}(\ell \in C_q(X)|Y = y)p_Y(y)p_{\tilde{Y}|Y}(\tilde{y}|y)\mathrm{d}y \\ &= \int M_q^c(\ell, y)p_{\tilde{Y}|Y}(\tilde{y}|y)\mathrm{d}y \\ &= \left[M_q^c(\ell, \cdot) * k(\cdot; \sigma^2)\right](\tilde{y}) \end{aligned} \tag{2}$$

其中第三个等式来自贝叶斯规则，第四个等式使用了 A.1，最后一个等式使用了 A.2，其中‘$*$’表示卷积运算。

利用上述结果，我们可以设计一个估计无噪声情况下 CP 阈值 $\hat{q}$ 的程序。具体来说，给定 q 的一个值，我们可以使用带噪声的校准数据来计算（近似）$M_q^n(\ell, y)$ 对于 $\ell, \tilde{y} \in \mathcal{Y}$。然后，(2) 表明为了获得每 $\ell \in \mathcal{Y}$ 的（近似）$M_q^c(\ell, \cdot)$，我们需要解决一个带有 $M_q^n(\ell, \cdot)$ 和噪声核 k 的一维反卷积问题。接下来，使用 (1) 中的关系，我们可以近似覆盖概率 $\mathbb{P}(Y \in C_q(X))$。如果它较大（相应地较小）于 $1 - \alpha$，我们需要稍微减小（相应地增大）q 并重复该过程。这种概念策略在算法 1 中给出，其中我们利用噪声模型具有色散性的事实，因此使用“噪声 CP”方法初始化 q 将导致一个较大的值，我们可以逐渐减小这个值直到达到停止准则 $\mathbb{P}(Y \in C_q(X)) \gtrapprox 1 - \alpha$。

Algorithm 1 计算 $\hat{q}$ 的概念性高层程序（详见第 3.2 节的实际实现）

Input: α，小 δ_q（默认：0.05），依赖阈值 q 的 $\mathcal{C}_q(\cdot)$ 预测区间规则，带有噪声标签的校准集 $\{x_i, \tilde{y}_i\}_{i=1}^n$，噪声核 $k(\cdot; \sigma^2)$。

初始化：使用 $\{x_i, \tilde{y}_i\}_{i=1}^n$ 对 $q \leftarrow$ 进行普通 CP 校准

repeat

$q \leftarrow q - \delta_q$

步骤 1：使用噪声标签计算 $M_q^n(\ell, \tilde{y})$

对于 $\ell, \tilde{y} \in \mathcal{Y}$:$M_q^n(\ell, \tilde{y}) \leftarrow \mathbb{P}(\ell \in C_q(X)|\tilde{Y} = \tilde{y})p_{\tilde{Y}}(\tilde{y})$

步骤 2：使用反卷积计算 $M_q^c(\ell, \tilde{y})$

对于 $\ell \in \mathcal{Y}$:$M_q^c(\ell, \cdot) \leftarrow$ 去卷积 $M_q^n(\ell, \cdot)$ 和 $k(\cdot; \sigma)$

步骤 3：计算覆盖概率

$\mathbb{P}(Y \in C_q(X)) \leftarrow \int M_q^c(\ell, \ell)\mathrm{d}\ell$

until $\mathbb{P}(Y \in C_q(X)) < 1 - \alpha$

返回 $\hat{q} \leftarrow q + \delta_q$

3.2. 实际实现

让我们解释在实践中如何实现算法 1 中的每一步。显然，主要的挑战是出现在 M_q^n 定义中的概率和分布是未知的，我们只能从有限样本 $\{x_i, \tilde{y}_i\}_{i=1}^n$ 中近似得到它们。这将需要使用离散化方法。然而，尽管我们的近似存在，我们在第 4 节将展示我们的方法具有很强的经验性能。

离散化。域 $\mathcal{Y}$ 通过将略大于 $[\min_i \tilde{y}_i, \max_i \tilde{y}_i]$ 的范围划分成宽度为 $\delta_y = 0.01$ 的区间 $\{B_\ell\}$ 来进行离散化。用 L 表示区间的数量。从现在起我们使用 $\ell, \tilde{y}$ 表示索引。每个带噪声的标签 $\tilde{y}_i$ 被分配到一个箱子里 B_ℓ，同时对应的输入 x_i 也被分配进去。

步骤 1。对于给定的 $q > 0$，我们通过如下计算得到一个 $L \times L$ 矩阵 $\hat{M}_q^n$ 来近似 M_q^n：$\hat{M}_q^n[\ell, \tilde{y}] = \frac{\sum_i \mathbb{I}\{B_\ell \subseteq C_q(x_i) \cap \tilde{y}_i \in B_{\tilde{y}}\}}{|\tilde{y}_i : \tilde{y}_i \in B_{\tilde{y}}|} \cdot \frac{1}{\delta_y} \frac{\sum_i \mathbb{I}\{\tilde{y}_i \in B_{\tilde{y}}\}}{n}$。也就是说，我们用其经验版本来近似概率。

步骤 2. 给定核的离散化，$\hat{k}$，我们通过求解凸优化问

题来近似 M_q^c

$$\min_{\substack{\hat{M}_q^c \in \mathbb{R}^{L\times L} \\ 0\le \hat{M}_q^c[\ell,\tilde{y}] \le 1/\delta_y}} \sum_{\ell} \left\| (\hat{M}_q^c[\ell,:] * \hat{k} - \hat{M}_q^n[\ell,:]) \odot \mathbf{1}_{\hat{M}_q^n[\ell,:]>\epsilon} \right\|_2^2 + \lambda \|\hat{M}_q^c\|_F^2$$

其中 $\odot$ 表示元素-wise 乘法，我们使用它来屏蔽 $\hat{M}_q^n[\ell,:]$ 中为空的 bin。我们发现这种屏蔽对于处理校准集是有限的事实是必要的。我们将正则化参数设置为 $\lambda = 0.01$。

步骤 3。我们使用 $\sum_\ell \hat{M}_q^c[\ell,\ell]\delta_y$ 作为覆盖率概率的近似值。

3.3. 估计标签噪声的水平

算法 1 需要了解噪声核 $k(\cdot;\sigma)$。特别是，在高斯噪声假设 $\tilde{Y}|Y \sim \mathcal{N}(0,\sigma^2)$ 下，方差 σ^2 需要被知晓或估计。让我们给出一个简单的估计器用于 σ^2，给定一个使用带有噪声标签训练的模型。未来的工作可以进一步研究估计标签噪声模型的问题。

我们从关系式开始

$$\begin{aligned} p_{\tilde{Y}|X}(\tilde{y}|x) &= \int p_{\tilde{Y}|Y,X}(\tilde{y}|y,x) p_{Y|X}(y|x) dy \\ &= \int p_{\tilde{Y}|Y}(\tilde{y}|y) p_{Y|X}(y|x) dy, \end{aligned}$$

其中第二个等号使用了 A.1。假设 $Y|X \sim \mathcal{N}(\mu(X), u^2(X))$ 连同 $\tilde{Y}|Y \sim \mathcal{N}(0,\sigma^2)$，积分实际上是高斯函数的卷积，给出了 $\tilde{Y}|X \sim \mathcal{N}(\mu(X), u^2(X)+\sigma^2)$。

由于模型 $(\hat{y}(\cdot), \hat{u}^2(\cdot))$ 是用高斯负对数似然训练的，对于样本 x，我们简单地通过 $\hat{u}^2(x)$ 来估计 $u^2(x)+\sigma^2$。因此，在具有 "简单" 样本 $\{x_i\}$ 的温和假设下，可以从训练集中 $\hat{u}^2(\cdot)$ 的最小值中估计 $u^2(x_i) \approx 0$，σ^2。在我们的实验中，我们观察到 $\hat{u}^2(\cdot)$ 的最低 1% 个值的平均值提供了令人满意的预测。

4. 实验

在本节中，我们评估我们的方法在保持覆盖率的同时减少置信区间的有效性。

数据集。我们使用了两个医学成像数据集。**骨龄** [21]：从 RSNA 儿科骨骼年龄数据集中获取的手部 CT 年龄回归。**胸部 X 光片** [22]：来自 NIH 的大规模胸部 X 光片集合，常用于疾病分类任务。这两个数据集的任务都是从图像中推断一个人的年龄。我们将数据分为 60% 训练集，30% 测试集和 10% 校准集。我们通过训练标签的均值和标准差（SD）对标签进行归一化。

噪声模型。我们在校准集（在第 4.2 节中也在训练集中）对标签归一化之前，通过添加高斯噪声扰动真实年龄，并将它们四舍五入到整数单位（对于 [21] 为月份，对于 [22] 为年份）来实现标签噪声。我们用 σ_{Retrue} 表示标签噪声的标准差乘以干净训练标签的经验标准差。例如，$\sigma_{Retrue} = 0.2$ 表示使用噪声水平为干净标签标准差的五分之一。

回归 DNN。我们的模型基于在 ImageNet 上预训练的 EfficientNet-B4 [23]。我们移除了最终的全连接层并添加了两个任务特定的头部：一个用于均值预测 $\hat{y}$，另一个用于对数方差 $\log \hat{u}^2$。我们使用高斯 NLL 损失训练模型。

比较的方法。我们比较了以下方法：(1) 甲骨文 CP，该方法使用干净的校准标签
(2) 噪声 CP，在带有噪声的校准集上应用标准符合预测（受 [15] 启发）；
(3) 我们的，我们的 CP 方法，如第 3 节中所提出的。对于我们使用的所有方法 $\alpha = 0.1$（如同文献中常见的那样），得分 $s(x,y) = |y-\hat{y}(x)|/\hat{u}(x)$ 及其决策规则 $\mathcal{C}_q(\cdot)$。我们报告了在 6 次试验中计算的 CP 阈值、区间长度和覆盖率百分比（针对干净的测试标签）的均值 (± 标准差)。

4.1. 校准中仅存在噪声标签

我们在干净的数据上训练了一个模型，并在带有由 $\sigma_{Retrue} = 0.2$ 生成的噪声标签的校准集上应用了 CP。我们使用 $\sigma = 0.2$ 应用我们的 CP 方法，但也使用其他值来评估其对该参数的鲁棒性。

结果报告在表 1 中。正如预期的那样，对于 Noisy CP，$\hat{q}$ 显著大于 Oracle CP 的情况。相应地，它受到较大的置信区间和过度覆盖的影响。相比之下，我们的方法通过正确的 σ 达到了接近于 Oracle 的表现，区间比 Noisy CP 小得多，同时提供了目标覆盖率。我们看到应用我们的方法时使用较小的 σ（可以理解为：

Table 1. 覆盖率和区间长度对于 α=0.1 和真实值 σ_{Retrue}=0.2。我们的方法也应用于“错误”的 σ 值。

数据集	方法	$\hat{q}$ 值	平均长度↓	覆盖率 (%)
胸部 X 光片	Oracle CP	2.48 ± 0.01	0.58 ± 0.00	90.02 ± 0.00
	Noisy CP	3.73 ± 0.01	0.95 ± 0.01	96.10 ± 0.30
	Ours w/ σ=0.15	2.71 ± 0.02	0.71 ± 0.02	91.45 ± 0.15
	Ours w/ σ=0.2	2.53 ± 0.02	0.67 ± 0.01	90.57 ± 0.22
	Ours w/ σ=0.25	2.36 ± 0.02	0.60 ± 0.01	88.52 ± 0.16
骨龄	Oracle CP	2.40 ± 0.02	0.66 ± 0.01	90.10 ± 0.00
	Noisy CP	3.31 ± 0.03	0.94 ± 0.01	94.14 ± 0.30
	Ours w/ σ=0.15	2.44 ± 0.02	0.79 ± 0.01	90.84 ± 0.15
	Ours w/ σ=0.2	2.42 ± 0.02	0.74 ± 0.02	90.11 ± 0.14
	Ours w/ σ=0.25	2.37 ± 0.02	0.60 ± 0.01	89.72 ± 0.17

Table 2. 覆盖率和区间长度对于胸部 X 光数据集，α=0.1 和 $\sigma_{Retrue} \in \{0.2, 0.3\}$。我们的方法应用于估计 $\hat{\sigma}$。

σ	方法	$\hat{q}$ 值	平均长度↓	覆盖率 (%)
σ_{Retrue}=0.2 $\hat{\sigma}$=0.205	Oracle CP	2.33 ± 0.02	0.66 ± 0.01	90.7 ± 0.21
	Noisy CP	3.27 ± 0.04	0.97 ± 0.01	95.7 ± 0.34
	Ours	2.33 ± 0.02	0.72 ± 0.01	90.0 ± 0.26
σ_{Retrue}=0.3 $\hat{\sigma}$=0.314	Oracle CP	2.50 ± 0.03	0.77 ± 0.01	90.4 ± 0.18
	Noisy CP	3.92 ± 0.05	1.29 ± 0.02	96.4 ± 0.38
	Ours	2.52 ± 0.02	0.84 ± 0.02	90.5 ± 0.32

不解决部分噪声）会导致过度覆盖和更大的区间（类似于 Noisy CP 与 Oracle CP 之间的关系），反之，如果使用较大的 σ 则相反。然而，对于我们的方法，这些指标的变化相对较小，这证明了它的鲁棒性。

4.2. 训练和校准中的噪声标签

我们考虑训练集和校准集中水平为 σ_{Retrue} 的标签噪声。我们使用第 3.3 节中描述的过程来获得 $\hat{\sigma}$，它是 σ_{Retrue} 的一个估计值，用于应用我们的 CP 方法。我们检查 $\sigma_{Retrue} \in \{0.2, 0.3\}$。

结果报告在表 2 中。为了成为一个有意义的性能界限，Oracle CP 也使用相同模型，这些模型是用噪声标签训练的（但有一个干净的校准集）。这解释了其性能与表 2 相比略有下降的原因，在 $\sigma_{Retrue} = 0.2$ 中也是如此。正如预期的那样，随着噪声水平的增加，Noisy CP 区间增大的情况变得更为显著。与上面展示的鲁棒性特性一致，我们的方法得益于 $\hat{\sigma}$ 接近 σ_{Retrue} 的事实，并表现出接近 Oracle 的性能，远优于 Noisy CP。

5. 结论

我们解决了在校准集中使用带噪声标签的 CP 应用于回归 DNNs 的问题。我们开发了一种基于迭代脱卷的估计无噪声 CP 阈值的过程，该过程基于坚实的数学推导。我们在两个带有高斯标签噪声的医学成像回归数据集上评估了该方法，并表明它优于现有的替代方法，并且性能接近于清洁标签设置。

6. REFERENCES

[1] Christian Leibig, Vaneeda Allken, Murat Seckin Ayhan, Philipp Berens, and Siegfried Wahl, “Leveraging uncertainty information from deep neural networks for disease detection,” Scientific Reports, vol. 7, no. 1, pp. 1–14, 2017.

[2] Sebastian Scher and Gabriele Messori, “Predicting weather forecast uncertainty with machine learning,” Quarterly Journal of the Royal Meteorological Society, vol. 144, no. 717, pp. 2830–2841, 2018.

[3] Ashwin Carvalho, Stéphanie Lefévre, Georg Schildbach, Jason Kong, and Francesco Borrelli, “Automated driving: The role of forecasts and uncertainty—a control perspective,” European Journal of Control, vol. 24, pp. 14–32, 2015.

[4] Netanell Avisdris, Leo Joskowicz, Brian Dromey, Anna L David, Donald M Peebles, Danail Stoyanov, Dafna Ben Bashat, and Sophia Bano, “Biometrynet: Landmark-based fetal biometry estimation from standard ultrasound planes,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022.

[5] Volodya Vovk, Alexander Gammerman, and Craig Saunders, “Machine-learning applications of algorithmic randomness,” in Proceedings of the Sixteenth International Conference on Machine Learning, 1999, pp. 444–453.

[6] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, Algorithmic learning in a random world, vol. 29, Springer, 2005.

[7] Anastasios N Angelopoulos, Stephen Bates, et al., “Conformal prediction: A gentle introduction,” Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494–591, 2023.

[8] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik, "Uncertainty sets for image classifiers using conformal prediction," in International Conference on Learning Representations, 2021.

[9] Lahav Dabah and Tom Tirer, "On temperature scaling and conformal prediction of deep classifiers," in Forty-second International Conference on Machine Learning, 2025.

[10] Charles Lu, Anastasios N Angelopoulos, and Stuart Pomerantz, "Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets," in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2022.

[11] Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer, "Fair conformal predictors for applications in medical imaging," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[12] Henrik Olsson, Kimmo Kartasalo, Nita Mulliqi, et al., "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," Nature Communications, vol. 13, no. 1, pp. 7761, 2022.

[13] Yuko Kato, David MJ Tax, and Marco Loog, "A review of nonconformity measures for conformal prediction in regression," Conformal and Probabilistic Prediction with Applications, pp. 369–383, 2023.

[14] Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng, "Conformal prediction: A data perspective," ACM Computing Surveys, 2025.

[15] Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano, "Label noise robustness of conformal prediction," Journal of Machine Learning Research, vol. 25, no. 328, pp. 1–66, 2024.

[16] Matteo Sesia, YX Rachel Wang, and Xin Tong, "Adaptive conformal classification with noisy labels," Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024.

[17] Coby Penso and Jacob Goldberger, "A conformal prediction score that is robust to label noise," in MICCAI Int. Workshop on Machine Learning in Medical Imaging (MLMI), 2024.

[18] Coby Penso, Jacob Goldberger, and Ethan Fetaya, "Conformal prediction of classifiers with many classes based on noisy labels," in Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications, 2025.

[19] Jase Clarkson, Wenkai Xu, Mihai i Cucuringu, and Gesine Reinert, "Split conformal prediction under data contamination," in Proceedings of the Symposium on Conformal and Probabilistic Prediction with Applications, 2024.

[20] Rotem Nizhar, Lior Frenkel, and Jacob Goldberger, "Clinical measurements with calibrated instance-dependent confidence interval," in Medical Imaging with Deep Learning, 2025.

[21] Safwan S. Halabi, William Prevedello, Michael Kalpathy-Cramer, and et al., "The rsna pediatric bone age machine learning challenge," Radiology, vol. 290, no. 2, pp. 498, 2019.

[22] National Institutes of Health (NIH), "Nih chest x-ray dataset," https://www.kaggle.com/datasets/nih-chest-xrays/data, 2017, Accessed: 2025-09-15.

[23] Mingxing Tan and Quoc V Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning. PMLR, 2019, pp. 6105–6114.