A Supplementary Material

In the following, we provide formal proofs, mathematical details, and extra figures. In addition, we supplement our submission with the implementation of our methods and experiments: https://github.com/mirkobunse/pkccn

A.1 Proof of Consistent CCN Learning

To establish the proof of Proposition 1, we first prove the following prerequisite, which goes back to optimal thresholds for accuracy in particular [27, supplementary material, proof of Lemma 7].

Proposition 2. *Assume binary CCN label noise according to Def. 1, where* $p_+ + p_- < 1$. *There exists a strictly monotone transformation, which does not depend on* x, *between the true clean probabilities* $\mathbb{P}(Y = +1 \mid X = x)$ *and the true noisy probabilities* $\mathbb{P}(\widehat{Y} = +1 \mid X = x)$. *Namely,* $\exists\, a > 0, b \in \mathbb{R}$, *such that,* $\forall\, x \in \mathcal{X} : \mathbb{P}(X = x) > 0$,

$$\mathbb{P}(\widehat{Y} = +1 \mid X = x) \;=\; a \cdot \mathbb{P}(Y = +1 \mid X = x) + b.$$

In particular, $a = 1 - p_+ - p_-$ *and* $b = p_-$.

Proof (Proposition 2). The following rearrangement employs i) the law of total probability and ii) the fact that any CCN noisy label $\widehat{y}$ only depends on the clean label y; more specifically, any $\widehat{y}$ is conditionally independent of the features x given y. Moreover, we employ iii) that in binary classification $\mathbb{P}(Y = -1 \mid X = x) = 1 - \mathbb{P}(Y = +1 \mid X = x)$. For all $x \in \mathcal{X} : \mathbb{P}(X = x) > 0$,

$$\begin{aligned}
&\mathbb{P}(\widehat{Y} = +1 \mid X = x) \\
&= \frac{1}{\mathbb{P}(X = x)} \mathbb{P}(\widehat{Y} = +1 \cap X = x) \\
&\overset{\text{i)}}{=} \frac{1}{\mathbb{P}(X = x)} \sum_{y \in \{+1,-1\}} \mathbb{P}(Y = y \cap \widehat{Y} = +1 \cap X = x) \\
&= \frac{1}{\mathbb{P}(X = x)} \sum_{y \in \{+1,-1\}} \mathbb{P}(\widehat{Y} = +1 \mid Y = y \cap X = x) \cdot \mathbb{P}(Y = y \cap X = x) \\
&\overset{\text{ii)}}{=} \frac{1}{\mathbb{P}(X = x)} \sum_{y \in \{+1,-1\}} \mathbb{P}(\widehat{Y} = +1 \mid Y = y) \cdot \mathbb{P}(Y = y \cap X = x) \\
&= \sum_{y \in \{+1,-1\}} \mathbb{P}(\widehat{Y} = +1 \mid Y = y) \cdot \mathbb{P}(Y = y \mid X = x) \\
&\overset{\text{iii)}}{=} (1 - p_+) \cdot \mathbb{P}(Y = +1 \mid X = x) \;+\; p_- \cdot \big(1 - \mathbb{P}(Y = +1 \mid X = x)\big) \\
&= a \cdot \mathbb{P}(Y = +1 \mid X = x) \;+\; b
\end{aligned}$$

Here, $a > 0$ due to the assumption $p_+ + p_- < 1$. □

Remark 4. Menon et al. [24, Proposition 5] and Scott et al. [30, Proposition 1] present similar one-to-one correspondences between noisy and clean prediction functions in the context of binary classification with mutually contaminated distributions. While this context might appear to be different from CCN learning, they further show [24, Eq. 3] that both settings are in fact equivalent. In particular, both settings allow us to learn a scoring function from noisily labeled data and optimize the decision threshold for predicting the clean classes.

We can now turn to the main proof of Proposition 1:

Proof (Proposition 1). Let $a = 1 - p_+ - p_- > 0$ and $b = p_-$. The following rearrangement employs i) Proposition 2 and ii) $m \to \infty$ with the fact that $\mathcal{A}$ is consistent. The Bayes-optimal classifier for $\mathcal{Q}$, with respect to the clean ground-truth labels, is

$$\begin{aligned} h^*(x) &= sign(\mathbb{P}(Y = +1 \mid X = x) - \theta^*) \\ &= sign(a \cdot \mathbb{P}(Y = +1 \mid X = x) + b - a \cdot \theta^* - b) \\ &\overset{\text{i)}}{=} sign(\mathbb{P}(\widehat{Y} = +1 \mid X = x) - (a \cdot \theta^* + b)) \\ &\overset{\text{ii)}}{=} sign(h_{\mathcal{A}}(x) - (a \cdot \theta^* + b)), \end{aligned}$$

which further tells us that the Bayes-optimal classifier for $\mathcal{Q}$, with respect to the CCN noisy labels, is

$$sign(h_{\mathcal{A}}(x) - \phi^*),$$

where $\phi^* = a \cdot \theta^* + b$ is the threshold that is optimal for the CCN noisy labels. Rearranging this equation for the clean-optimal threshold

$$\theta^* \;=\; \frac{\phi^* - b}{a},$$

which we intend to find, yields the claim of Proposition 1. □

A.2 Proof of the Hypothesis Test for PK-CCN Learnability

Before establishing the proof of Theorem 1, we note that our choice of setting $\omega = \frac{p_-}{1-p_-}$ is a direct consequence of observing that $p_- = \frac{1}{R+1}$ and $\omega = \frac{1}{R}$:

$$\begin{aligned} p_- \;=\; \frac{1}{R+1} \quad &\Rightarrow \quad R \;=\; \frac{1}{p_-} - 1 \;=\; \frac{1-p_-}{p_-} \\ &\Rightarrow \quad \omega \;=\; \frac{1}{R} \;=\; \frac{p_-}{1-p_-} \end{aligned}$$

Proof (Theorem 1). We emphasize that Theorem 1 defines $N_{\widehat{y}}$ as the number of clean positives with a noisy label $\widehat{y} \in \{+1, -1\}$. Hence, we argue about true positives instead of the predicted positives that Def. 2 originally describes.

What remains to be shown for proving Theorem 1 is that the null hypotheses of the two hypothesis tests,

$$h_0^{\text{Def.2}} : \ \lambda_+ \ \leq \ \omega \cdot \lambda_-,$$
$$h_0^{\text{Th.1}} : \ p_+ + p_- \ \geq \ 1$$

are indeed equivalent. To this end, let N be the number of true positives in the training set, so that the Poisson rates in $h_0^{\text{Def.2}}$ are $\lambda_+ = (1-p_+)N$ and $\lambda_- = p_+ N$. Rearranging $h_0^{\text{Th.1}}$,

$$\begin{aligned}
& p_+ + p_- && \geq \ 1 \\
\Leftrightarrow \quad & p_+ p_- && \geq \ 1 - p_- - p_+ + p_+ p_- \\
& && = \ (1-p_-)(1-p_+) \\
\Leftrightarrow \quad & p_+ \frac{p_-}{1-p_-} && \geq \ 1 - p_+ \\
\Leftrightarrow \quad & \omega p_+ N && \geq \ (1-p_+)N \\
\Leftrightarrow \quad & \omega \lambda_- && \geq \ \lambda_+,
\end{aligned}$$

yields the equivalence of $h_0^{\text{Def.2}}$ and $h_0^{\text{Th.1}}$. □

A.3 F_1 Score Estimation in Spite of CCN

Menon et al. [24] estimate the clean F_1 score as

$$F_1(\theta;\, h, p_-, p_+) \ = \ \frac{2 \cdot \pi \cdot \text{tpr}(\theta; p_-, p_+)}{\pi(1 + \text{tpr}(\theta; p_-, p_+)) + (1-\pi)(1 - \text{tnr}(\theta; p_-, p_+))}, \tag{4}$$

where the true positive rate and the true negative rate, with respect to the clean ground-truth labels, are estimated as

$$\begin{aligned}
\text{tpr}(\theta; p_-, p_+) \ &= \ 1 - \frac{a(1 - \widehat{\text{tnr}}(\theta)) + (1-b)(1 - \widehat{\text{tpr}}(\theta)) - a}{1 - a - b}, \\
\text{tnr}(\theta; p_-, p_+) \ &= \ 1 - \frac{(1-a)(1 - \widehat{\text{tnr}}(\theta)) + b(1 - \widehat{\text{tpr}}(\theta)) - b}{1 - a - b}.
\end{aligned} \tag{5}$$

Here, $\widehat{\text{tpr}}(\theta) = \mathbb{P}(h(x) > \theta \mid \widehat{Y} = +1)$ and $\widehat{\text{tnr}}(\theta) = \mathbb{P}(h(x) \leq \theta \mid \widehat{Y} = -1)$ are the rates of true positives and true negatives with respect to the noisy labels and $\widehat{\pi} = \mathbb{P}(\widehat{Y} = +1)$ is the noisy class prior. These rates can be estimated from noisily labeled data alone. Moreover, let

$$a \ = \ \frac{p_-(1 - p_+ - \widehat{\pi})}{\widehat{\pi}(1 - p_+ - p_-)}, \quad b \ = \ \frac{p_+(\widehat{\pi} - p_-)}{(1-\widehat{\pi})(1 - p_+ - p_-)}, \quad \pi \ = \ \frac{\widehat{\pi} - p_-}{1 - p_+ - p_-}. \tag{6}$$

A.4 CD Diagram in Terms of the f_ω Score

Fig. 4 displays CD diagrams that compare performances in terms of the F_1 score. We now complete this picture with Fig. 5, which follows an identical setup but replaces the F_1 score with the f_ω score as a performance measure.

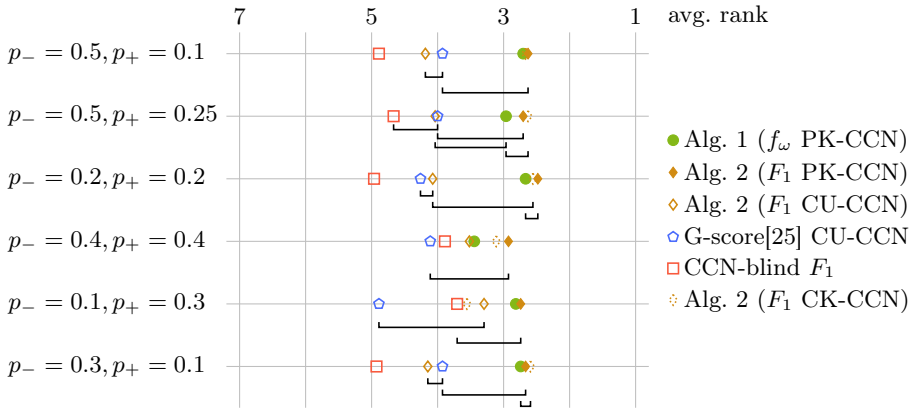

Fig. 5. This overview of our results is analogous to Fig. 4, but employs the f_ω score instead of the F_1 score to determine the performance of each method.

In comparison to Fig. 4, Fig. 5 produces larger groups of indistinguishable methods. While Fig. 5 does not single out one clear winner, it still conveys that PK-CCN methods have a clear advantage over CU-CCN methods.