Neural Conditional Density Estimation for Regression Tasks - Practical Work in AI

Alexander Krauck
Department of Machine Learning
Johannes Kepler University Linz
Upper Austria, Austria
alexander.krauck@gmail.com

1 Introduction

In order to properly define the goal of conformal prediction, some definitions are needed. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Furthermore, let $(\mathbf{X}_i, \mathbf{Y}_i)$ be a sequence of random variables on the index set $\mathcal{I}$ with $\mathbf{X_i} : \Omega \to \mathbb{R}^n$ and $\mathbf{Y_i} : \Omega \to \mathbb{R}^m$ for all i. Moreover, let all $(\mathbf{X}_i, \mathbf{Y}_i)$ be exchangable, which means that for any permutation π of the index set $\mathcal{I}$ the join probability distribution remains the same, i.e.

$$\forall \mathbf{x}_i \in \mathbb{R}^n, \mathbf{y}_i \in \mathbb{R}^m : \mathbb{P}\left(\bigcap_{i \in \mathcal{I}} (X_{\pi(i)} \leq \mathbf{x}_i, Y_{\pi(i)} \leq \mathbf{y}_i)\right) = \mathbb{P}\left(\bigcap_{i \in \mathcal{I}} (X_i \leq \mathbf{x}_i, Y_i \leq \mathbf{y}_i)\right)$$

Furthermore, let $I : \Omega \to \mathcal{I}$ be a random variable defined on the probability space that selects an index at random with the further simplification that every index has the same probability to be selected (uniform).

Moreover, let $a \in (0, 1)$ be some significance level. Then the goal of conformal prediction is to find a function $U : \mathbb{R}^n \to \mathcal{B}(\mathbb{R}^m)$ that can predict the subsets with the smallest Lebesgue measure λ marginalized over $\mathcal{I}$ with significance a. That means we want $\mathbb{P}(\mathbf{Y}_I \in U(\mathbf{X}_I)) = a$ with $\int_\Omega \lambda(U(\mathbf{X}_{I(\omega)}(\omega)))d\mathbb{P}(\omega)$ small.

Formally that means we want

$$\min_U \quad \int_{\mathbb{R}^{m+n}} p(\mathbf{z})\lambda(U(\mathbf{x}))d\mathbf{z} \tag{1}$$

$$\text{s.t.} \quad \int_{\mathbb{R}^{m+n}} \mathbb{1}_{\mathbf{y} \in U(\mathbf{x})} d\mathbf{z} = a \tag{2}$$

To be clear, $p(\mathbf{z})$ simply is the probability density function of and event $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \sim (\mathbf{X}_I, \mathbf{Y}_I)$ which is defined by

$$p(\mathbf{x}, \mathbf{y}) := \frac{d^2\mathbb{P}(\mathbf{X}_I \leq \mathbf{x}, \mathbf{Y}_I \leq \mathbf{y})}{d\mathbf{x}d\mathbf{y}}$$

The marginal and conditional probability density functions can then be found by integrating out and normalizing with the marginal respectively.

If U is being calculated by using first Conditional Density Estimation (CDE) and then using Highest Density Regions (HDRs) as defined by [Hyndman, 1996] to obtain a significance level of s, then U is a function of the CDE method and the significance level s. HDR is by [Hyndman, 1996] defined as:

$$H(f_a) = \{\mathbf{y} : f(\mathbf{y}) \geq f_a\}$$

with

$$f_a = \max_{f_a} \left\{ f_a \in \mathbb{R}^+ : \mathbb{P}\left(\mathbf{y} \in H(f_a)\right) \geq a \right\}$$

but can also be written as:

$$H\left(f, a\right) = \left\{ \mathbf{y} \in \mathbb{R}^m : f(\mathbf{y}) \geq \max_b \left\{ b \in \mathbb{R}^+ : \mathbb{P}\left(\left\{ \hat{\mathbf{y}} \in \mathbb{R}^m : f(\hat{\mathbf{y}}) \geq b \right\}\right) \geq a \right\} \right\} \tag{3}$$

where f is an arbitrary probability density function (PDF) and a is the credibility level. Note that $\mathbb{P}$ here is different from the one defined in the beginning and only here for defining HDRs. As the CDE method in my case is parametric like mixture density networks it is more reasonable to write U a function of the weights and the significance (and implicitly also the architecture of the CDE of course). So considering that, the initial goal of conformal prediction can be rewritten as:

$$\min_{\boldsymbol{\Theta}} \quad \int_{\mathbb{R}^{m+n}} p(\mathbf{z}) \lambda(H(p(\hat{\mathbf{y}} \mid \mathbf{x}; \boldsymbol{\Theta}), a)) d\mathbf{z} \tag{4}$$

$$\text{s.t.} \quad \int_{\mathbb{R}^{m+n}} \mathbb{1}_{\mathbf{y} \in H(p(\hat{\mathbf{y}} \mid \mathbf{x}; \boldsymbol{\Theta}), a)} d\mathbf{z} = a \tag{5}$$

where it is important that the $\hat{\mathbf{y}}$ is not the one we integrate over but more a demonstrative artefact we write to denote that $p(\hat{\mathbf{y}} \mid \mathbf{x}; \boldsymbol{\Theta})$ is a conditional density estimator.

In the following if we write $p(\hat{\mathbf{y}} \mid \mathbf{x}; \boldsymbol{\Theta})$ we mean a conditional density estimator parameterized with $\boldsymbol{\Theta}$ which implicitly contains the architecture and the weights of the method and if we write $p(\hat{\mathbf{y}} \mid \mathbf{x})$, the true conditional density is meant. Notice, that the goal of this optimization problem is to optimize w.r.t. $\boldsymbol{\Theta}$ as it is basically the component that completely defines $p(\hat{\mathbf{y}} \mid \mathbf{x}; \boldsymbol{\Theta})$. This in turn means we need to find the argmin of the optimization above.

Let $\boldsymbol{\Theta}^*$ be the argmin of the equation above, we would like to show that

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} \int_{\mathbb{R}^{m+n}} p(\mathbf{z}) \log p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\Theta}) d\mathbf{z}$$

which is the maximum likelihood estimator (MLE). This is of great importance if we wish to optimize a CDE method w.r.t. the maximum likelihood objective function in order to implicitly optimize for the conformal prediction objective function when using HDRs.

In order to show this powerful statement we first need to develop some insight into the workings of the components in conformal prediction.

Lemma 1.1. *Let $p : \mathbb{R}^m \to \mathbb{R}$ be a continuous density function. Then, the function $\lambda_{H_p}(a) := \lambda(H(p, a))$, representing the size of the highest density region (HDR) for PDF p with coverage a, is convex, i.e.,*

$$\frac{\partial^2 \lambda_{H_p}(a)}{\partial a^2} \geq 0.$$

Still need backup for that here: Furthermore, for density functions p that are not continuous, the concept of convexity is generalized to average convexity. Specifically, $\lambda_{H_p}(a)$ is considered convex if there exists a finite $k < \infty$ such that, over any interval $[a_1, a_2]$ with $a_2 - a_1 < k$, the function's average rate of increase does not decrease. This captures the macroscopic behavior of the HDR size expansion despite local discontinuities, ensuring that the concept is applied within a practical range that accommodates the inherent variability of p. This is left without proof.

Proof. Without loss of generality, assume $a_1 < a_2$ with both $a_1, a_2 \in (0, 1)$. For any $\alpha \in (0, 1)$, let $a := \alpha a_1 + (1 - \alpha) a_2$. By the definition of H_p, the set $H(p, a)$ encompasses points up to the highest densities corresponding to coverage a.

This implies that $\lambda_{H_p}(a_1) \leq \lambda_{H_p}(a) \leq \lambda_{H_p}(a_2)$. Define $k = a - a_1$, hence $\lambda_{H_p}(a) = \lambda_{H_p}(a_1 + k)$, which necessitates $H(p, a)$ to include densities for a coverage level of $a_1 + k$, unlike $H(p, a_1)$ that covers up to a_1. Consequently, $H(p, a_1) \subset H(p, a)$, indicating that to transition from $H(p, a_1)$ to $H(p, a)$, only points with density less than

$$b_1 = \max_b \left\{ b \in \mathbb{R}^+ : \mathbb{P}\left(\{\hat{\mathbf{y}} \in \mathbb{R}^m : p(\hat{\mathbf{y}}) \geq b\}\right) \geq a_1 \right\}$$

can be utilized. In contrast, to achieve $H(p, a) \subset H(p, a_2)$, points must have a density lower than

$$b_2 = \max_b \left\{ b \in \mathbb{R}^+ : \mathbb{P}\left(\{\hat{\mathbf{y}} \in \mathbb{R}^m : p(\hat{\mathbf{y}}) \geq b\}\right) \geq a \right\}.$$

Given that $b_1 \geq b_2$, which suggests that as coverage increases, we are forced to incorporate points of decreasing density, this relationship establishes an indirect proportionality between the density thresholds (b_1, b_2) and the HDR size expansion. This implies that as we increase the coverage (from a_1 to a_2), the incremental increase in HDR size (from $\lambda_{H_p}(a_1)$ to $\lambda_{H_p}(a_2)$) becomes progressively more steep, indicative of a non-negative second derivative $\frac{\partial^2 \lambda_{H_p}(a)}{\partial a^2} \geq 0$, and thus, confirming the convexity of λ_{H_p} with respect to coverage a. □

Lemma 1.2. *For any probability density function $f : \mathbb{R}^m \to \mathbb{R}$ and $p : \mathbb{R}^m \to \mathbb{R}$ with the same coverage size under significance a, that is,*

$$\lambda(H(f, a)) = \lambda(H(p, a))$$

it holds that if we measure the coverage of those HDRs with samples drawn w.r.t. p, that the coverage measured with $H(p, a)$ will be higher and the coverage level w.r.t. $H(p, a)$ will be greater-equal to a. Formally:

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(f,a)} d\mathbf{y} \leq \int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(p,a)} d\mathbf{y} \geq a$$

Proof. We split $H(p, a)$ and $H(f, a)$ into subsets: $H(f, a) = A \cup B$ and $H(f, a) = A \cup C$ with $H(f, a) \cap H(p, a) = A$, $H(f, a) \setminus A = B$ and $H(p, a) \setminus A = C$.

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(f,a)} d\mathbf{y} = \int_{H(f,a)} p(\mathbf{y}) d\mathbf{y} = \int_A p(\mathbf{y}) d\mathbf{y} + \int_B p(\mathbf{y}) d\mathbf{y}$$

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(p,a)} d\mathbf{y} = \int_{H(p,a)} p(\mathbf{y}) d\mathbf{y} = \int_A p(\mathbf{y}) d\mathbf{y} + \int_C p(\mathbf{y}) d\mathbf{y}$$

since the A part of the integrals is equal we can ignore it for comparing $H(f, a)$ and $H(p, a)$ coverage. So, we need to show:

$$\int_B p(\mathbf{y}) d\mathbf{y} \leq \int_C p(\mathbf{y}) d\mathbf{y}$$

This can be shown by proofing $\forall \mathbf{y}_B \in B \forall \mathbf{y}_C \in C : p(\mathbf{y}_B) \leq p(\mathbf{y}_C)$ because $\lambda(B) = \lambda(C)$. So, let $\mathbf{y}_B$ and $\mathbf{y}_C$ be arbitrary from the corresponding sets. Then we know that $\mathbf{y}_B \in H(p, a)$, which means that $p(\mathbf{y}_C) \geq b$ where $b = \max_b \left\{ b \in \mathbb{R}^+ : \mathbb{P}\left(\{\hat{\mathbf{y}} \in \mathbb{R}^m : p(\hat{\mathbf{y}}) \geq b\}\right) \geq a \right\}$ is the maximum density bound such that the coverage level of a is still given w.r.t. p. However, since $\mathbf{y}_C \notin H(p, a)$ this means that $p(\mathbf{y}_C) < b$, which shows the first part of the proof.

The second part of the proof is to show that

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(p,a)} d\mathbf{y} \geq a$$

which follows easily by the definition of HDI. □

In order to finally be able to use this lemma efficiently we need to define something like an inverse of the HDR w.r.t. the significance, which based on the input $\mathbf{x}$ and the length $\lambda(H(p(\mathbf{y} \mid \mathbf{y}), a))$ gives us the significance level that we would need to insert in H together with the true distribution at x to obtain the same size of the distribution but with the above lemma always has a larger or equal coverage.

Definition 1.1 (Inverse Maximum Likelihood HDR)**.** Let $f, p : \mathbb{R}^m \to \mathbb{R}$ be two probability density functions and let $a \in (0, 1)$ be some coverage level. Then we define the inverse Maximum Likelihood HDR as:

$$IH(f, a) = \underset{b \in (0,1)}{\arg\max}\, \lambda(H(p, b)) \leq \lambda(H(f, a))$$

Note that it is not guaranteed that $\lambda(H(p, b)) = \lambda(H(f, a))$ if p is not continuous. Using this definition together with Lemma 1.2 establishes a very powerful tool to make proofs related to HDR.

Now we will proof the marginal equivalent to Lemma 1.2 that will be essentially for showing the final result.

Lemma 1.3. *For any probability density function $f : \mathbb{R}^m \to \mathbb{R}$ and $p : \mathbb{R}^m \to \mathbb{R}$, where p is continuous. with the same coverage size, in average using p as the underlying probability distribution, under significance a, that is,*

$$\int_{R^m} p(\mathbf{y}) \lambda(H(f, a)) d\mathbf{y} = \int_{R^m} p(\mathbf{y}) \lambda(H(p, a)) d\mathbf{y}$$

it holds that if we measure the coverage of those HDRs with samples drawn w.r.t. p, that the coverage measured with $H(p, a)$.Formally:

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(f,a)} d\mathbf{y} \leq \int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(p,a)} d\mathbf{y}$$

Proof. For this proof we require everything that we have shown so far as it is a very powerful statement already. First, lets write down the formula for the coverage of f:

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(f,a)} d\mathbf{y}$$

In order to show this statement we require to form an upper bound this this expression for which we can show the statement. This upper bound can conveniently be built using the inverse maximum likelihood HDR. It holds that:

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(f,a)} d\mathbf{y} \leq \int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(p, IH(f,a))} d\mathbf{y}$$

This statements holds for any arbitrary f that we can define. Note, that if we transform $H(f, a)$ to $H(p, IH(f, a))$. If we do not assume continuity of p, then we need to also consider that

$$\int_{R^m} p(\mathbf{y}) \lambda(H(f, a)) d\mathbf{y} \geq \int_{R^m} p(\mathbf{y}) \lambda(H(p, IH(f, a))) d\mathbf{y}$$

as a property of the IH which means the coverage size might become smaller. However, as p also works with the same distribution and has coverage $\bar{\lambda}$ actually the above will always be an equality.

Using this we can now consider that:

$$\int_{\mathbb{R}^m} p(\mathbf{y}) \mathbb{1}_{\mathbf{y} \in H(p, IH(f,a))} d\mathbf{y} = \int_{\mathbb{R}^m} p(\mathbf{y}) IH(f, a) d\mathbf{y} = \mathbb{E}\left[IH(f, a)\right]$$

where the mid equality holds because of how IH is defined. Note that we don't know yet if $\mathbb{E}\left[IH(f,a)\right] = a$ or not. We also consider that

$$\int_{R^m} p(\mathbf{y})\lambda(H(p, IH(f,a)))d\mathbf{y} = \mathbb{E}\left[\lambda(H(p, IH(f,a)))\right]$$

.

Now we can apply convexity of $\lambda(H(p,a))$ to see that:

$$\bar{\lambda} = \mathbb{E}\left[\bar{\lambda}\right] = \mathbb{E}\left[\mathbb{E}\left[\lambda(H(p, IH(f,a)))\right]\right] \geq \mathbb{E}\left[\lambda(H(p, \mathbb{E}\left[IH(f,a)\right]))\right]$$

where we can apply a double expectation since $\bar{\lambda}$ is a constant. However, we know that $\int_{\mathbb{R}^m} p(\mathbf{y})\mathbb{1}_{\mathbf{y}\in H(p,a)}d\mathbf{y} = \int_{\mathbb{R}^m} p(\mathbf{y})\mathbb{1}_{\mathbf{y}\in H(p,a)}d\mathbf{y} = a$ and thus that:

$$\bar{\lambda} = \mathbb{E}\left[\lambda(H(p,a))\right]$$

which implies that

$$a \geq \mathbb{E}\left[IH(f,a)\right]$$

since a smaller coverage size implies smaller significance level on the same distribution. That is what we needed to show.

□

Now we have all tools in order to proof the wanted statement.

Theorem 1.4 (MLL is equivalent with Optimal Conformal Prediction)**.** *We want to show that:*

$$\arg\max_{\mathbf{\Theta}} \int_{\mathbb{R}^{m+n}} p(\mathbf{z}) \log p(\mathbf{y} \mid \mathbf{x}; \mathbf{\Theta}) d\mathbf{z}$$

equals

$$\arg\min_{\mathbf{\Theta}} \quad \int_{\mathbb{R}^{m+n}} p(\mathbf{z})\lambda(H(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a))d\mathbf{z} \tag{6}$$

$$\text{s.t.} \quad \int_{\mathbb{R}^{m+n}} \mathbb{1}_{\mathbf{y}\in H(p(\hat{\mathbf{y}}\mid\mathbf{x};\mathbf{\Theta}),a)} d\mathbf{z} = a \tag{7}$$

with the variables defined as in the very beginning. Moreover we for now assume the $p(\mathbf{y} \mid \mathbf{x})$ is continuous always and that $\forall \mathbf{x} \in \mathbb{R}^n : \max_{\mathbf{\Theta}} p(\mathbf{y} \mid \mathbf{x}; \mathbf{\Theta}) = p(\mathbf{y} \mid \mathbf{x})$.

Proof. For brevity, let $H_{\mathbf{\Theta}} := H(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a)$ and $H_{\mathbf{\Theta}^*} := H(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}^*), a)$.

In this case let's assume that now we have some $\mathbf{\Theta} \neq \mathbf{\Theta}^*$. We can show that if for this $\mathbf{\Theta}$ it holds that if

$$\int_{\mathbb{R}^{m+n}} p(\mathbf{z})\lambda(H_{\mathbf{\Theta}})d\mathbf{z} < \int_{\mathbb{R}^{m+n}} p(\mathbf{z})\lambda(H_{\mathbf{\Theta}^*})d\mathbf{z}$$

it implies

$$\int_{\mathbb{R}^{m+n}} p(\mathbf{z})\mathbb{1}_{\mathbf{y}\in H_{\mathbf{\Theta}}} d\mathbf{z} < a$$

which would finish the proof, since it'd show that for any parameter set $\mathbf{\Theta}$, that produces a smaller average HDR than the the MLE, the constraint would be violated and it is clear that the MLE would fulfill the constraint since it is clearly calibrated.

We can upper bound the coverage with the IH:

$$\int_{\mathbb{R}^{m+n}} p(\mathbf{z})\mathbb{1}_{\mathbf{y}\in H(p(\hat{\mathbf{y}}\mid\mathbf{x};\mathbf{\Theta}),a)} d\mathbf{z} \leq \int_{\mathbb{R}^{m+n}} p(\mathbf{z})\mathbb{1}_{\mathbf{y}\in H(p(\hat{\mathbf{y}}\mid\mathbf{x}), IH(p(\hat{\mathbf{y}}\mid\mathbf{x};\mathbf{\Theta}),a))} d\mathbf{z}$$

$$= \mathbb{E}\left[IH(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a)\right]$$

We know that there exists some $k \in \mathbb{R}^+$ such that:

$$\mathbb{E}\left[\lambda(H_{\mathbf{\Theta}^*})\right] - k = \mathbb{E}\left[\lambda(H_{\mathbf{\Theta}})\right] = \mathbb{E}\left[\lambda(H(p(\hat{\mathbf{y}} \mid \mathbf{x}), IH(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a)))\right]$$

Due to convexity it now holds:

$$\mathbb{E}\left[\mathbb{E}\left[\lambda(H(p(\hat{\mathbf{y}} \mid \mathbf{x}), IH(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a)))\right]\right] \geq \mathbb{E}\left[\lambda(H(p(\hat{\mathbf{y}} \mid \mathbf{x}), \mathbb{E}\left[IH(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a)\right]))\right]$$

$$\mathbb{E}\left[\lambda(H_{\mathbf{\Theta}^*})\right] = \mathbb{E}\left[\lambda(H(p(\hat{\mathbf{y}} \mid \mathbf{x}), a)))\right] > \mathbb{E}\left[\lambda(H(p(\hat{\mathbf{y}} \mid \mathbf{x}), \mathbb{E}\left[IH(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a)\right]))\right]$$

which implies that $\mathbb{E}\left[IH(p(\hat{\mathbf{y}} \mid \mathbf{x}; \mathbf{\Theta}), a)\right] < a$ which was to be shown.

□

This now shows our desired result, that in fact, by maximizing the likelihood of a CDE model to afterwards apply HDR to obtain small but calibrated regions is equivalent to directly trying to minimize the size of the HDR while maintaining calibration.

When assuming that our CDE models can reasonably well approximate the true underlying distribution this is a powerful result that allows us to focus fully on optimizing the likelihood for CDE and we get optimal conformal prediction for free.

References

R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2): 120–126, 1996.