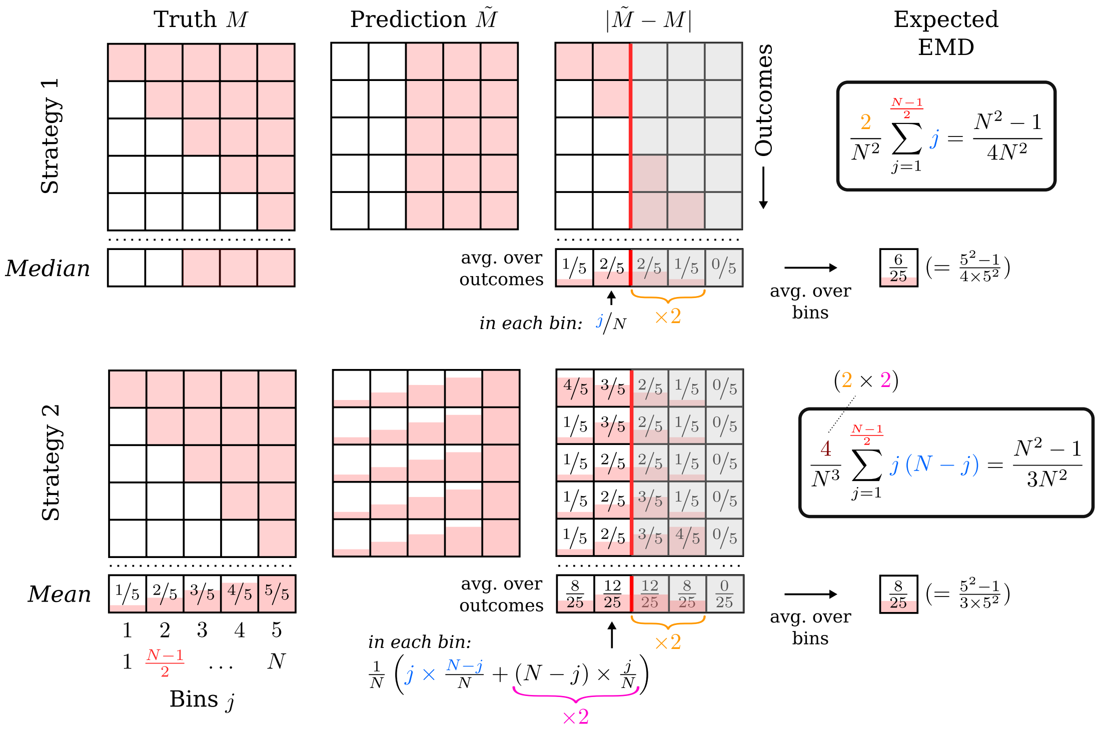
SUPPLEMENTARY MATERIAL

The Earth Mover's Pinball Loss: Quantiles for Histogram-Valued Regression

Florian List

S1. Expected EMD in the Toy Example for a Single Draw



Truth M
Prediction $\tilde{M}$
$|\tilde{M} - M|$
Expected EMD
Strategy 1
Outcomes
$\frac{2}{N^2}\sum_{j=1}^{\frac{N-1}{2}} j = \frac{N^2-1}{4N^2}$
Median
avg. over outcomes
1/5 2/5 2/5 1/5 0/5
in each bin: j/N
×2
avg. over bins
$\frac{6}{25}$ $(= \frac{5^2-1}{4\times 5^2})$
Strategy 2
(2 × 2)
$\frac{4}{N^3}\sum_{j=1}^{\frac{N-1}{2}} j\,(N-j) = \frac{N^2-1}{3N^2}$
Mean
1/5 2/5 3/5 4/5 5/5
$\frac{8}{25}$ $\frac{12}{25}$ $\frac{12}{25}$ $\frac{8}{25}$ $\frac{0}{25}$
$\frac{8}{25}$ $(= \frac{5^2-1}{3\times 5^2})$
1 2 3 4 5
1 $\frac{N-1}{2}$ … N
Bins j
in each bin:
$\frac{1}{N}\left(j \times \frac{N-j}{N} + (N-j) \times \frac{j}{N}\right)$
×2

Figure S1. Sketch illustrating the calculation of the expected EMD with Strategies 1 and 2, where the predicted cumulative histogram $\tilde{M}$ is given by the *median* and the *mean* over all possible cumulative histograms M, respectively. We show the case $N = 5$ as considered in the main body. Within each grid, the columns correspond to the bins $j \in \{1, \ldots, N\}$, and each row belongs to a possible outcome $Y \in \{1, \ldots, N\}$. The filling of each square indicates the value, from 0 (empty) to 1 (completely filled). The rightmost grids show the absolute difference $|\tilde{M} - M|$ between the predicted and true cumulative histogram for each outcome and bin. The expected EMD for each strategy is given by the mean of $|\tilde{M} - M|$ over the outcomes (vertically) and over the bins (horizontally), and can therefore be read off as the filled area fraction (e.g., 6 out of 25 squares are filled with Strategy 1 for $N = 5$). Due to the symmetry of $|\tilde{M} - M|$ w.r.t. the central bin, the total filled areas to the right and to the left of the red vertical line are equal, for which reason it is sufficient to consider $j \in \{1, \ldots, (N-1)/2\}$ and to multiply the result by two (in yellow) in the calculation of the expected EMD. With Strategy 1, averaging $|\tilde{M}_j - M_j|$ over the outcomes gives j/N for bins $j \leq (N-1)/2$, whereas one obtains $1/N\big(j(N-j)/N + (N-j)j/N\big)$ with Strategy 2, for example $\frac{12}{25} = \frac{1}{5}\,(2 \times {}^3\!/_5 + 3 \times {}^2\!/_5)$ for $j = 2$ and $N = 5$.

In this section, we briefly revisit the toy example from the main body (Sec. 3.1) and discuss the minimization of the expected EMD for a single draw. Recall that the experiment consists of x times randomly drawing a numbered ball from $\mathcal{N} = \{1, \ldots, N\}$ with replacement, where each draw is described by a random variable $Y \sim \mathcal{U}\{1, N\}$ that follows a discrete uniform distribution. The histogram label $m = (m_j)_{j=1}^N$ in bin j expresses the fraction of times the number j

was drawn (that is, $m_j = 0$ if number j was never drawn, and $m_j = 1$ if j was drawn x times). For a *single* draw, i.e. $x = 1$, we noted that a histogram estimate that places all probability mass in the central bin minimizes the expected EMD towards the true outcome. Therefore, this is what a NN trained using the EMD (and hence an EMPL-trained NN just as well for $\tau = 0.5$) predicts in this case, as can be seen in the top panel in Fig. 2. Here, we calculate the expected EMD as a function of N for this strategy and compare it with the case of a histogram estimate that uniformly distributes the probability mass over the bins. We restrict ourselves to the case of odd N, such that there is a unique central bin. In the discrete case of normalized histograms in 1D, the EMD (or 1-Wasserstein distance) between predicted histogram $\tilde{m}$ and observed realization m is given by

$$W_1(\tilde{m}, m) = \frac{1}{N} \sum_{j=1}^{N} |\tilde{M}_j - M_j|, \tag{S1}$$

where $M_j = \sum_{r=1}^{j} m_r$ denotes the observed cumulative histogram, and similarly for $\tilde{M}_j$. Note that in order not to overload notation, we do not distinguish between the random variables $m = (m_j)_{j=1}^N$ (and $M = (M_j)_{j=1}^N$) that express the random histogram values in each bin and realizations of these random variables in this Supplementary Material.

Strategy 1 (median): *Putting everything on* $(N + 1)/2$
In this strategy, the estimated histogram $\tilde{m}^1 = (\tilde{m}_j^1)_{j=1}^N$ is $\tilde{m}_j^1 = 1$ for $j = (N + 1)/2$ (the central bin) and $\tilde{m}_j^1 = 0$ otherwise. This choice minimizes the expected EMD, which immediately follows from the fact that the associated *cumulative* histogram $\tilde{M}_j^1 = 0$ for $j < (N+1)/2$ and $\tilde{M}_j^1 = 1$ for $j \geq (N+1)/2$ is the *median* over the possible cumulative outcomes in each bin (see the top left grid in Fig. S1), recalling that the median minimizes the mean absolute error. To obtain the expected EMD, we will first compute the mean of $|\tilde{M}_j^1 - M_j^1|$ over the N possible outcomes ($Y = 1, \ldots, N$) for each bin j, and then take the mean over the bins. Since the possible outcomes as well as the estimated histogram are symmetric w.r.t. the central bin, it is sufficient to consider $j \in \{1, \ldots, (N - 1)/2\}$ (i.e., the bins to the left of the red vertical line in the rightmost grids in Fig. S1), for which one finds that the expected absolute difference between the estimated and true cumulative histogram is given by

$$\mathbb{E}_m \left[|\tilde{M}_j^1 - M_j| \right] = \frac{j}{N}. \tag{S2}$$

Accounting for the symmetry and averaging over the bins $j = 1, \ldots, N$ yields

$$\mathbb{E}_m \left[W_1(\tilde{m}^1, m) \right] = \frac{2}{N^2} \sum_{j=1}^{\frac{N-1}{2}} j = \frac{N^2 - 1}{4N^2}. \tag{S3}$$

Strategy 2 (mean): *Uniformly distributing the probability mass*
For comparison, we consider another possible strategy, where the probability mass is uniformly distributed over the bins, and the estimated histogram $\tilde{m}^2 = (\tilde{m}_j^2)_{j=1}^N$ is defined by $\tilde{m}_j^2 = 1/N$ (and hence $\tilde{M}_j^2 = j/N$). Note that this choice corresponds to the *mean* over all possible realizations of the cumulative histogram. As above for Strategy 1, we compute the expected absolute difference between $\tilde{M}_j^2$ and M_j in each bin $j \in \{1, \ldots, (N - 1)/2\}$, which now yields

$$\mathbb{E}_m \left[|\tilde{M}_j^2 - M_j| \right] = \frac{1}{N} \left(j \frac{(N - j)}{N} + (N - j) \frac{j}{N} \right) = \frac{2\, j\, (N - j)}{N^2}, \tag{S4}$$

observing that there are j outcomes with $|\tilde{M}_j^2 - M_j| = (N - j)/N$ and $(N - j)$ outcomes with $|\tilde{M}_j^2 - M_j| = j/N$ (see the bottom right grid in Fig. S1). Exploiting the symmetry about the central bin again and averaging over the bins, we obtain

$$\mathbb{E}_m \left[W_1(\tilde{m}^2, m) \right] = \frac{4}{N^3} \sum_{j=1}^{\frac{N-1}{2}} j\, (N - j) = \frac{N^2 - 1}{3N^2}. \tag{S5}$$

Comparing the two strategies, we find that

$$\mathbb{E}_m \left[W_1(\tilde{m}^2, m) \right] = \frac{4}{3}\, \mathbb{E}_m \left[W_1(\tilde{m}^1, m) \right], \tag{S6}$$

which confirms that predicting the median over all possible cumulative histogram realizations in each bin (Strategy 1) indeed leads to a smaller expected EMD as compared to the mean (Strategy 2) for $N > 1$.

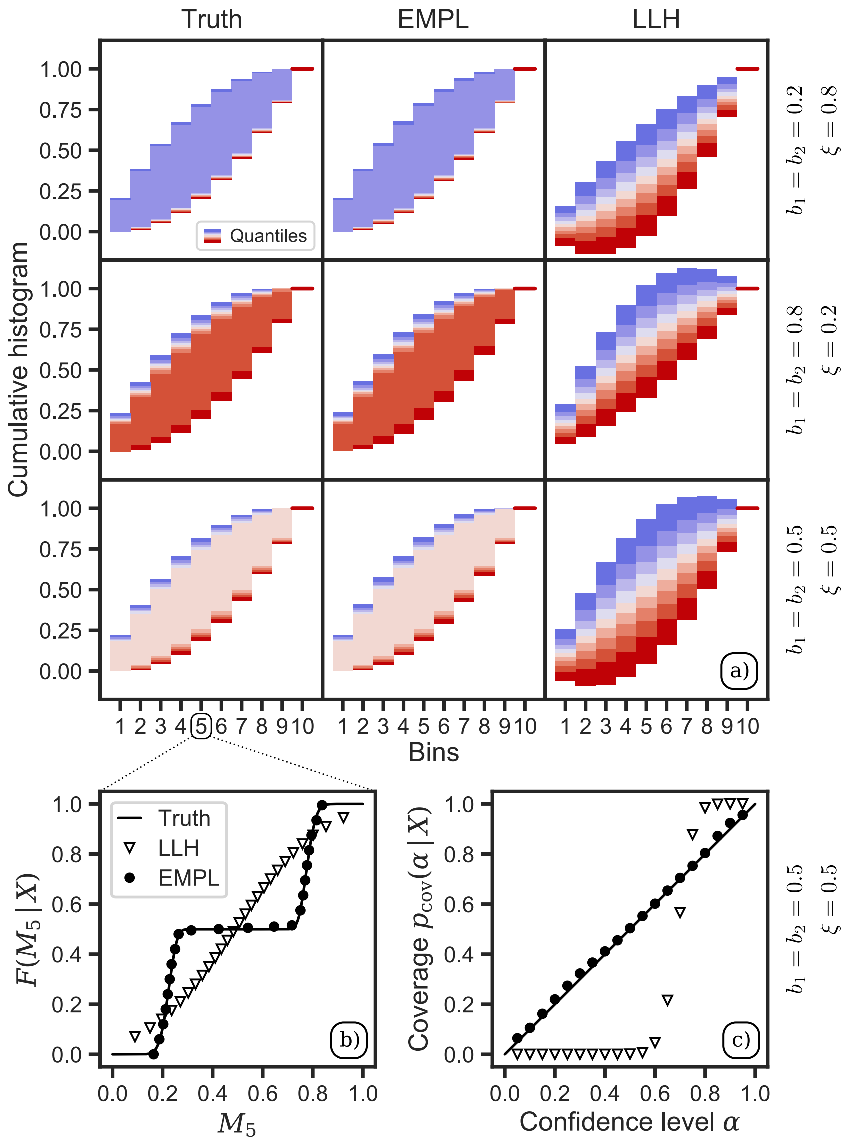


Figure S2. Panel a): True and estimated distributions of the cumulative histograms, for three different inputs $X = (b_1, b_2, \xi)$ as specified next to each row. The colored regions show the $5 - 95\%$ quantiles in steps of 10%. The EMPL predictions (center) closely resemble the truth (left), whereas a NN trained to maximize the Gaussian log-likelihood within each bin (right) is clearly not able to capture the bimodal distribution. Panel b): CDF of the cumulative histogram value M_j in bin $j = 5$ for the input $X = (0.5, 0.5, 0.5)$ considered in the bottom row of panel a). Panel c): Calibration plot for the same input $X = (0.5, 0.5, 0.5)$. The bin-averaged coverage probability $p_{\text{cov}}(\alpha \,|\, X)$ is given by the fraction of samples and bins that fall within the symmetric α-interquantile range around the median. Perfectly calibrated uncertainties would lie on the identity line.

S2. A Bimodal Toy Example

In the main body, we have presented the results of EMPL-trained NNs for three problems in different areas. In this section, we consider an additional toy example where the distribution of the cumulative histogram within each bin is *bimodal*. The 3-dimensional NN input $X = (b_1, b_2, \xi)$ determines the width of each mode ($b_1, b_2 \in [0, 1]$) and the probability for the histogram label to follow the first mode ($\xi \in [0, 1]$; the probability to follow the second mode is thus $1 - \xi$). For the specific parameterization that we use to generate the histogram labels, as well as for the implementation details, we refer to our Github repository. We use $N = 10$ histogram bins and train a MLP with 2 hidden layers consisting of 256 neurons each by minimizing the EMPL for 10,000 batch iterations at batch size 2,048. As in the football example, we compare our EMPL results with a NN trained by maximizing a Gaussian log-likelihood for the CDF in each bin.

Panel a) in Fig. S2 shows the true and estimated distributions of the cumulative histograms for three different NN inputs X, namely for narrow modes and a high probability for the first mode ($b_1 = b_2 = 0.2$, $\xi = 0.8$; top row), for wide modes and a high probability for the second mode ($b_1 = b_2 = 0.8$, $\xi = 0.2$; middle row), and for the symmetric case with moderately wide modes ($b_1 = b_2 = \xi = 0.5$; bottom row). The distribution predicted by the EMPL-trained NN closely resembles the truth for all considered inputs, whereas a Gaussian likelihood yields completely unsatisfactory results for these highly non-Gaussian uncertainties. For the symmetric case in the bottom row ($X = (0.5, 0.5, 0.5)$), we also plot the CDF in bin 5, see panel b). The EMPL prediction agrees with the truth, while a Gaussian CDF clearly cannot account for the two modes of the distribution. Finally, we consider the calibration of the uncertainties by computing the bin-averaged coverage probability $p_{\text{cov}}(\alpha \,|\, X)$ as a function of the confidence level α, conditional on the input $X = (0.5, 0.5, 0.5)$. We calculate the coverage probability as the fraction of samples and bins for which the true value falls within the α-interquantile range symmetrically around the median. We use 65,536 histogram realizations for the evaluation, and we only average over those bins where the true cumulative histogram lies within $[\varepsilon, 1 - \varepsilon]$ for $\varepsilon = 10^{-5}$ in order to avoid biases due to numerical inaccuracies far below the magnitudes of interest. We confirmed that the results are not sensitive to the exact choice of ε. With the EMPL, the maximum deviation from perfect calibration ($p_{\text{cov}}(\alpha \,|\, X) = \alpha$) is $< 3\%$ for all the considered confidence levels α, in contrast to $> 50\%$ with a bin-wise Gaussian log-likelihood loss function. This example demonstrates that the EMPL is able to accurately recover complex posterior distributions over plausible histograms, conditional on an input vector X.

S3. Tensorflow Implementation

We provide a basic `Tensorflow` implementation of the EMPL below (tested with `Tensorflow 2.3` and `Python 3.8`), which can be adapted according to the specific use case. Note that the input to the function is given by the true and estimated *density* histograms (not the cumulative histograms), which should be properly normalized, e.g. using a softmax activation function (however, the normalization is not checked by the function below).

```python
import tensorflow as tf
def empl(m, m_tilde, tau=0.5, alpha=0.0, scope="empl"):
    """
        Computes the Earth Mover's Pinball Loss:
        :param m:            true histogram labels (shape: n_batch x n_bins)
        :param m_tilde:      histogram predictions (shape: n_batch x n_bins)
        :param tau:          quantile level tau in [0, 1]
        :param alpha:        smoothing parameter alpha (>= 0)
        :param scope:        scope name
        :returns             Earth Mover's Pinball Loss between the density histograms
                             m and m_tilde for the quantile level tau
    """
    assert len(m.shape) == len(m_tilde.shape) == 2, "Only 2D tensors are supported!"
    assert m.shape[0] == m_tilde.shape[0], "Batch dimensions do not agree!"
    assert m.shape[1] == m_tilde.shape[1], "Bin dimensions do not agree!"

    with tf.name_scope(scope):
        # Density histograms -> cumulative histograms
        M = tf.cumsum(m, axis=1)
        M_tilde = tf.cumsum(m_tilde, axis=1)

        # Compute difference
        delta = M_tilde - M

        # Non-smooth loss (default)
        if alpha == 0.0:
            mask = tf.cast(tf.greater_equal(delta, tf.zeros_like(delta)), delta.dtype) - tau
            loss = mask * delta

        # Smooth loss
        else:
            loss = -tau * delta + alpha * tf.math.softplus(delta / alpha)

        # Return mean over bins and batch
        return tf.reduce_mean(loss)
```