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

Florian List [1]

Abstract

Although ubiquitous in the sciences, histogram data have not received much attention by the Deep Learning community. Whilst regression and classification tasks for scalar and vector data are routinely solved by neural networks, a principled approach for estimating histogram labels as a function of an input vector or image is lacking in the literature. We present a dedicated method for Deep Learning-based histogram regression, which incorporates cross-bin information and yields *distributions* over possible histograms, expressed by τ-quantiles of the cumulative histogram in each bin. The crux of our approach is a new loss function obtained by applying the pinball loss to the cumulative histogram, which for 1D histograms reduces to the Earth Mover's distance (EMD) in the special case of the median ($\tau = 0.5$), and generalizes it to arbitrary quantiles. We validate our method with an illustrative toy example, a football-related task, and an astrophysical computer vision problem. We show that with our loss function, the accuracy of the predicted *median* histograms is very similar to the standard EMD case (and higher than for per-bin loss functions such as cross-entropy), while the predictions become much more informative at almost no additional computational cost. ○

1. Introduction

Histograms, i.e. approximate representations of the distribution of numerical data obtained by binning the data into adjacent, non-overlapping bins, are frequently used across the disciplines. Typical examples are precipitation histograms in meteorology (e.g. Nicholls et al. 1997), population pyramids in demographics and ecology (e.g. Weeks 2020, and

Figure 1. Comic of our method for Deep Learning-based histogram regression: a neural network (NN) f_θ with weights θ is trained to estimate arbitrary τ-quantiles of the *cumulative* histogram $(M_j)_{j=1}^N$ in each bin $j \in \{1, \dots N\}$ (treated as N random variables), conditional on an input X. That is, the j-th NN output $\tilde{Q}_j^\theta(X; \tau) \approx Q_{M_j}(\tau \mid X)$ is an approximation of the true quantile function of M_j given X (i.e., the pre-image of τ under the CDF $F(M_j \mid X)$, see panel $\star$). Thus, the entire *distribution* of possible histograms is learned as a function of X. The values of the associated *density* histogram $(\tilde{q}_j)_{j=1}^N$, defined as $\tilde{q}_j = \tilde{Q}_j - \tilde{Q}_{j-1}$, increase early (late) for high (low) quantile levels τ.

color histograms in photography and image processing (e.g. Novak & Shafer 1992). While the invention of histograms is commonly attributed to Karl Pearson (Pearson, 1895), the use of bars for the representation of data can be traced back to the Middle Ages (Oresme, 1486). In Deep Learning, histograms appear in different contexts: they can be used as neural network (NN) *inputs* (e.g. Saadl et al. 2009; Rebetez et al. 2016), different variants of *hidden* histogram layers have been proposed (Wang et al., 2016; Sedighi & Fridrich, 2017; Peeples et al., 2020), block-wise histograms have been employed for feature pooling in Chan et al. (2015), and histogram loss functions were introduced in Ustinova & Lempitsky (2016); Zholus & Putin (2020). Furthermore, histograms of the trainable NN weights in different layers can shed light on whether the NN training is progressing properly.

In contrast, the task of regressing histogram *labels* based on an input vector (or image) X using NNs has not received much attention to date. This is despite the great potential of Deep Learning for identifying complex and nonlinear

---

[1]The University of Sydney, Sydney Institute for Astronomy, School of Physics, A28, NSW 2006, Australia. Correspondence to: Florian List <florian.list@sydney.edu.au>.
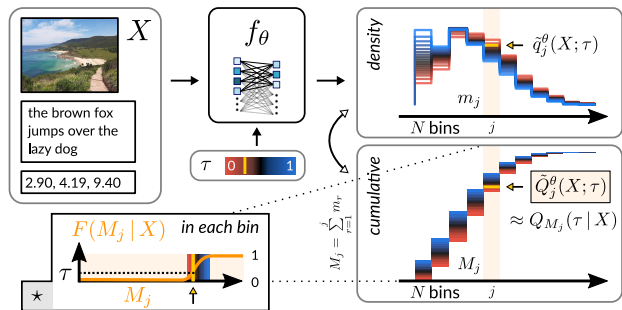
relations between an input X and an associated histogram, which is a common problem in many areas. For instance, Bellerby (2007) used a NN to predict rainfall histograms based on satellite-derived input data, Liu et al. (2020) considered the Deep Learning-based estimation of dose-volume histograms for radiotherapy planning, and Sharma et al. (2020) presented a CNN for the recovery of object size histograms from images, taking fly larvae and breast cancer cell data as examples.

Clearly, an ad-hoc approach is to treat each bin separately and to use a standard *regression* loss function such as the l^1 or l^2 error (mean absolute error and mean squared error, respectively), or a standard *classification* loss (softmax activation + cross-entropy loss); the latter assuming that the histograms sum up to unity (with the true label vector components lying anywhere in $[0, 1]$ instead of $\{0, 1\}$ as in the case of one-hot coded class labels for an actual classification problem). However, a drawback of these approaches is that the inherent ordering of the histogram bins is ignored, and cross-bin correlations are thus disregarded. Such an ordering may also be present in classification tasks: namely, whenever a continuous variable is discretized into bins, such as for image-based age estimation with labels "child", "adult", and "senior", as opposed to viewing the problem as a regression task, where the age is estimated as a number (e.g. in years). For these scenarios, Hou et al. (2016) suggested the use of the (squared) Earth Mover's distance (EMD; Rubner et al. 2000) as a loss function. The EMD measures the minimal amount of work needed to transform a distribution into another, and therefore penalizes the NN more when placing probability mass into bins far from the correct one, whereas the cross-entropy loss considers each bin in isolation. Another cross-bin loss function with a similar motivation is the Cumulative Jenssen–Shannon divergence (Nguyen & Vreeken, 2015; Jin et al., 2018). The important difference between ordered classification and our setting is, however, that we are interested in the *entire histogram*, not only in the argmax, which becomes the estimated class label for ordered classification, while the remaining estimated class probabilities are typically discarded. Therefore, we need the NN to correctly predict the value (or even the entire distribution of potential values) *in every bin*. This bears similarity to concepts such as Label Distribution Learning (LDL; Geng 2016), where the entire label distribution is relevant. However, LDL does not assume an underlying ordering of the labels, and categorical labels (e.g. "sky", "plant", "mountain") are supported, whereas our approach is specifically tailored to histograms.

In this work, we introduce a method for the NN-based estimation of conditional histograms. Since each input X can potentially belong to an entire *distribution* of output histograms, regressing a single "mean histogram" is often not sufficient, especially in applications where the *range* of

possible values in each bin for a specific input may have severe implications as in e.g. medicine. For this reason, we base our approach in *quantile regression* (Koenker & Bassett, 1978), enabling us to estimate arbitrary quantiles of the cumulative histogram in each bin. Specifically, we naturally extend the EMD for 1D distributions by allowing for asymmetry, in analogy to the pinball loss being an asymmetric generalization of the l^1 loss. In Sec. 2, we introduce the EMD and the pinball loss, and we define the *Earth Mover's Pinball Loss* (EMPL) by combining the two. Then, we demonstrate the effectiveness of our method in three scenarios in Sec. 3: first, we consider a toy example that can be phrased in terms of drawing numbered balls from an urn. Second, we consider an application in sports and use the EMPL to estimate league table positions. Finally, we consider a problem from γ-ray astronomy: the recovery of the brightness distributions of point-sources from photon-count maps. We conclude this work in Sec. 4.

2. Earth Mover's Pinball Loss

In this section, we introduce the EMPL for the task of histogram-valued quantile regression. We start by formally defining the optimization problem to be solved. Then, we recall the definitions of the EMD and the pinball loss, and proceed by defining the EMPL as a natural asymmetric extension of the EMD, which allows us to obtain quantiles for cumulative distribution functions (and hence for cumulative histograms in the discrete setting).

2.1. Problem formulation

We consider the task of learning a mapping from an independent (random) variable X to a corresponding distribution over output histograms with $N \in \mathbb{N}$ bins. We express this distribution over output histograms in terms of their quantiles, which has the advantage that no closed form for the distribution needs to be specified, making our method suitable for highly non-Gaussian and multimodal distributions. Recall that for a real-valued random variable Y with a strictly monotonic CDF $F_Y(y) = P(Y \leq y)$ and $\tau \in (0, 1)$, the τ-th quantile of Y is defined as

$$Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}. \quad (1)$$

For the sake of simplicity, we assume that the histogram values $(m_j)_{j=1}^N$ are normalized, i.e. $m_j \in [0, 1]$ and $\sum_{j=1}^N m_j = 1$, but our approach can easily be extended to arbitrary (non-negative) histograms by appending the total histogram count before normalization as an additional NN output. Further, we define the *cumulative* histogram by setting $M_j = \sum_{r=1}^j m_r$, and we write $M = (M_j)_{j=1}^N$.

Let f_θ be a NN with trainable weights θ, whose task is to predict τ-quantiles of the cumulative histogram $\tilde{Q}^\theta(X; \tau)$.

Our goal is to determine optimal parameters θ^* such that

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{X,\tau} \left[\left\| Q_M(\tau \mid X) - \tilde{Q}^\theta(X;\tau) \right\|_1 \right], \quad (2)$$

where $Q_M(\tau \mid X)$ is the vector-valued function that gathers the true τ-quantiles of the cumulative histogram M from all the bins $j = 1, \ldots, N$, given X. The expected value is taken over the input $X \sim P_X$ and uniformly over the quantile levels $\tau \sim U(0,1)$, and $\| \cdot \|_1$ is the l^1-norm on $\mathbb{R}^N$ that sums up the approximation errors from all the bins to a single number. A sketch of the histogram regression process is shown in Fig. 1.

2.2. Earth Mover's distance

As a first step towards solving Eq. (2), we introduce the EMD (Rubner et al., 2000), which is a distance measure between probability distributions rooted in the Optimal Transportation problem (Villani, 2009). As will be seen later, minimizing the EMD between the true and estimated histograms yields NN weights θ such that $\tilde{Q}^\theta(X;\tau) \approx Q_M(\tau \mid X)$ for the specific case of the *median* ($\tau = 0.5$).

Intuitively, the EMD measures the minimal amount of work that needs to be done in order to turn the "pile of dirt" (or earth) given by the PDF of distribution u into that of another distribution v. The definition of the EMD is in terms of "signatures", defined as sets of clusters each of which contains a certain amount of mass, and permits different total masses for different signatures. In the case of equal masses, however, it can be shown (Levina & Bickel, 2001) that the EMD is equivalent to the Wasserstein distance (Villani, 2003; Arjovsky et al., 2017).

Interpreting the normalized histograms $(m_j)_{j=1}^N$ as discretizations of continuous PDFs, we directly introduce the EMD in the continuous framework of the Wasserstein distance, which formally reads as follows:

For $d \in \mathbb{N}$, $p \in [1, \infty)$, let u, v be Borel probability measures on $\mathbb{R}^d$ with finite p-moments. Then, the p-Wasserstein distance is defined as (Villani, 2009)

$$W_p(u,v) = \left(\inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, d\pi(x,y) \right)^{1/p}, \quad (3)$$

where $\Gamma(u,v)$ is the collection of joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals u and v for the first and second argument, respectively. The definition of the p-Wasserstein distance does *not* require the measures u and v to be absolutely continuous w.r.t. the Lebesgue measure λ and is equally well-defined for discrete measures such as the Dirac measure, in which case the aforementioned notion of discrete clusters containing points can be recovered.

In this work, we restrict ourselves to the 1-Wasserstein distance in the one-dimensional case, i.e. $p = 1$ and $d = 1$,

in which the otherwise difficult calculation of the Wasserstein distance is greatly simplified and admits the following closed-form solution (Ramdas et al., 2017):

$$W_1(u,v) = \int_{\mathbb{R}} |U(t) - V(t)| \, dt, \quad (4)$$

where U and V are the CDFs of u and v, respectively, implying that the 1-Wasserstein distance is simply given as the L^1-distance between the CDFs of the two distributions in the 1D case, which arises from a notion of monotonicity that the optimal transport plan needs to satisfy (see Ramdas et al. 2017).

2.3. Pinball loss

Now, we turn towards the problem of quantile regression. For a scalar random variable Y, let $\tilde{y}$ be an approximation of the true quantile function $Q_Y(\tau)$. A suitable distance for comparing $\tilde{y}$ with observed values y is the pinball loss function (Fox & Rubin, 1964; Koenker & Bassett, 1978; Koenker & Hallock, 2001; Ferguson, 2014), defined as

$$\mathcal{L}_\tau^{\text{pin}}(y, \tilde{y}) = (y - \tilde{y}) \left(\tau - \mathbb{I}_{(y < \tilde{y})} \right)$$
$$= \begin{cases} \tau(y - \tilde{y}), & \text{if } y \geq \tilde{y}, \\ (\tau - 1)(y - \tilde{y}), & \text{if } y < \tilde{y}. \end{cases} \quad (5)$$

The pinball loss is constructed in such a way that its expectation is minimized by $\tilde{y} = Q_Y(\tau)$, which follows immediately from setting the derivative of the expected loss function w.r.t. $\tilde{y}$

$$\frac{\partial \mathbb{E}_Y[\mathcal{L}_\tau^{\text{pin}}(Y, \tilde{y})]}{\partial \tilde{y}} = (1 - \tau) F_Y(\tilde{y}) - \tau(1 - F_Y(\tilde{y}))$$
$$= F_Y(\tilde{y}) - \tau, \quad (6)$$

to zero, which yields the minimum at $\tilde{y} = Q_Y(\tau)$.

2.4. Quantile regression for histograms

Having introduced the EMD between probability distributions and the pinball loss for quantile regression, we now combine the two for the task of *histogram-valued regression*. We proceed in the continuous framework and subsequently consider the discrete case (i.e., histograms instead of PDFs).

For probability measures u and v, we define the EMPL as

$$\mathcal{L}_\tau(u,v) = \int_{\mathbb{R}} (U - V) \left(\tau - \mathbb{I}_{(U < V)} \right) \, dt, \quad (7)$$

where U and V are again the CDFs of u and v, respectively, and we suppress the argument t for brevity. This loss function can be viewed as an asymmetric extension of Eq. (4) in the spirit of the pinball loss in Eq. (5), with the asymmetry governed by the quantile level of interest τ. We can

decompose the integral into two regions and write

$$\mathcal{L}_\tau(u,v) = (1-\tau)\int_{U<V}|U-V|\,dt + \tau\int_{U\geq V}|U-V|\,dt,\tag{8}$$

from which the following bounds in terms of the 1-Wasserstein distance follow immediately:

$$\eta_- W_1(u,v) \leq \mathcal{L}_\tau(u,v) \leq \eta_+ W_1(u,v) \leq W_1(u,v),\tag{9}$$

where $\eta_- = \min\{\tau, 1-\tau\}$ and $\eta_+ = \max\{\tau, 1-\tau\}$. Note in particular that for the median ($\tau = 0.5$), one obtains

$$\mathcal{L}_{0.5}(u,v) = \frac{1}{2}\int_{\mathbb{R}}|U-V|\,dt = \frac{1}{2}W_1(u,v),\tag{10}$$

and the 1-Wasserstein distance in the case $d=1$ is recovered up to the factor of $1/2$ (see Eq. (4)). For $\tau \neq 0.5$, the EMPL is not symmetric and generally $\mathcal{L}_\tau(u,v) \neq \mathcal{L}_\tau(v,u)$, but rather $\mathcal{L}_\tau(u,v) = \mathcal{L}_{1-\tau}(v,u)$. Figuratively speaking, one could think of moving probability mass up or down a hill whose slope is determined by the quantile level of interest, making it more difficult to move probability mass upwards than downwards.

In the discrete setting, Eq. (7) becomes

$$\mathcal{L}_\tau(u,v) = \frac{1}{N}\sum_{j=1}^{N}\left[(U_j - V_j)\left(\tau - \mathbb{I}_{(U_j < V_j)}\right)\right],\tag{11}$$

where $U_j = \sum_{r=1}^{j}u_r$ and similarly for V_j. We remark that the EMPL as defined in Eq. (11) implicitly assumes the "distance" d_{ij} between two bins i and j in the notion of "work" when moving probability mass to be proportional to the distance between the bin indices, i.e. $d_{ij} \propto |i-j|$. For example, if one uses equally spaced (logarithmically spaced) bins in terms of the underlying variable R whose distribution is described by the histogram, the distance scales linearly with R (with $\log R$).

Coming back to the problem formulation, it now becomes apparent that training a NN using the EMPL as defined in Eq. (11) provides an (approximate) solution to Eq. (2):

Proposition 1. *For each fixed input $X = x$ and quantile level $\tau \in (0,1)$, a NN returning the conditional quantiles of the cumulative histogram, i.e. $\tilde{Q}^\theta(x;\tau) = Q_M(\tau\,|\,x)$, minimizes the expected τ-EMPL between observed cumulative histograms $\bar{M}(x)$ and the NN prediction $\tilde{Q}^\theta(x;\tau)$.*

Proof. This follows directly from plugging $U = \bar{M}(x)$ and $V = \tilde{Q}^\theta(x;\tau)$ into Eq. (11) and using the same argument as in Eq. (6) for each bin $j \in \{1, \ldots N\}$. □

An in-depth theoretical (convergence) analysis of the EMPL is beyond the scope of this paper and left to future work; however, our results in Sec. 3 are encouraging and confirm its suitability for diverse practical use cases.

2.5. Implementation details

The expectation over the inputs X in Eq. (2) is approximated as usual by training the NN on a large number of representative training samples $(X_s)_{s=1}^{S}$ with $X_s \sim P_X$ by means of a mini-batch gradient descent method. As for the expectation over the quantile level τ, we follow Tagasovska & Lopez-Paz (2019) and draw an individual quantile level τ for each input X_s from a uniform distribution $\tau \sim U(0,1)$ during the NN training. Compared with NNs that are trained for a single quantile level τ, the authors of that work showed that simultaneously estimating all the quantile levels greatly reduces *quantile crossing* in the case of scalar-valued NNs. The quantile levels τ appear at two places in the NN: 1) they are fed as an additional NN input in order for the NN to know which quantile level shall be estimated, and 2) they are used in the computation of the loss.

In practice, we obtain $\tilde{Q}$ by 1) estimating N logits $(\tilde{l}_j)_{j=1}^{N}$ (one per bin), 2) applying the softmax function $\tilde{q}_j = \text{softmax}(\tilde{l})_j$, which yields a normalized *density* histogram $(\tilde{q}_j)_{j=1}^{N}$, and 3) setting $\tilde{Q}_j = \sum_{r=1}^{j}\tilde{q}_r$, which enforces $\tilde{Q}_N = 1$. This implies that the NN prediction is properly normalized for all τ; moreover, monotonicity of the predicted cumulative histograms for each fixed τ is guaranteed because of $\tilde{Q}_j = \tilde{Q}_{j-1} + \tilde{q}_j$ with $\tilde{q}_j \in (0,1)$. The monotonicity of the quantiles *within* each bin is not strictly enforced, but it is encouraged by Eq. (11). Although we rarely ever encountered quantile crossing in our experiments once the NN is trained, crossing penalty terms as proposed by Takeuchi et al. (2006) could be incorporated in our framework without difficulty.

We also consider a Smoothed EMPL that is differentiable everywhere, derived by replacing the pinball loss by the smooth approximation proposed in Zheng (2011) (and applied to NNs in Hatalis et al. 2019), which yields

$$\mathcal{L}_\tau^\alpha(u,v) = \frac{1}{N}\sum_{j=1}^{N}\left[\tau\left(U_j - V_j\right)\right.$$
$$\left. + \alpha\log\left(1 + \exp\left(\frac{V_j - U_j}{\alpha}\right)\right)\right],\tag{12}$$

where $\alpha > 0$ is a smoothing parameter. In the limit $\alpha \searrow 0$, $\mathcal{L}_\tau(u,v)$ is recovered, and for any $\alpha > 0$, $\mathcal{L}_\tau^\alpha(u,v) - \mathcal{L}_\tau(u,v)$ is τ-independent, as follows immediately from the definitions. Note that $\log(1 + \exp(\cdot))$ is the softplus function, which is readily available in most machine learning libraries. An alternative approach that also provides differentiability everywhere is to consider an "L^2-version" of the EMPL, leading to the estimation of τ-expectiles rather than τ-quantiles (Aigner et al., 1976; Newey & Powell, 1987), which generalize the *mean* instead of the median and occasionally find use in financial risk estimation, but lack intuitive interpretability.
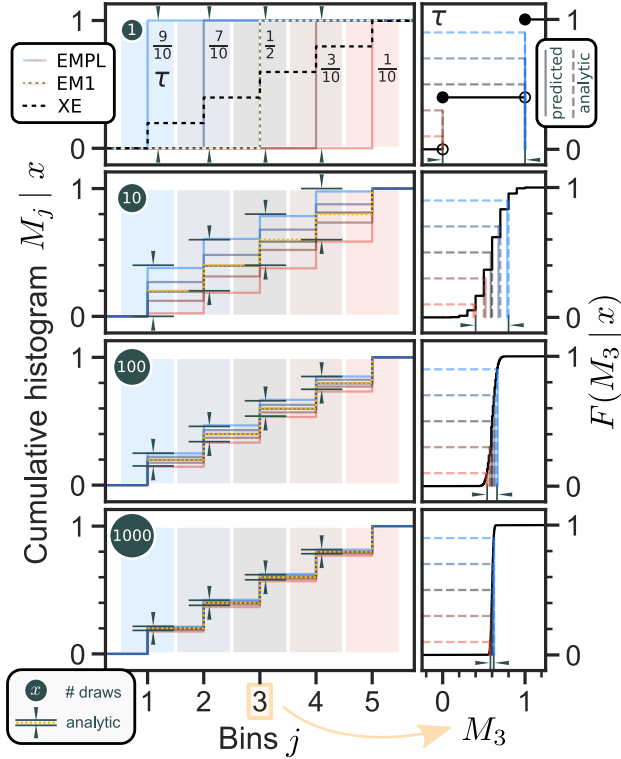
3. Results

Figure 2. *Left*: EMPL predictions for the toy example, for $x = 1$, 10, 100, and 1,000 draws. Each colored line corresponds to the MLP prediction for a particular quantile level τ (1/10, ..., 9/10), specified next to it in the top panel. The gray markers delimit the true inter-quantile range between the lowest and highest considered quantile levels, and the true median is indicated by the golden dotted line. For a single draw $x = 1$, the predictions of NNs trained using the 1-Wasserstein distance (EM1) and the cross-entropy loss (XE; computed w.r.t. to the density histogram $(m_j)_{j=1}^N$) are also shown: both produce a correct "average histogram" in a certain sense (namely median and mean, respectively), but a single number per bin is not sufficient to properly reflect the *distribution* of possible histograms. *Right*: Analytic CDF of the cumulative histogram M_3 in the central bin $j = 3$ (solid black line). The predicted quantiles (solid vertical lines) agree with the analytic quantiles (dashed vertical lines) for all values of x and quantile levels τ (dashed horizontal lines). As $x \to \infty$, all the quantiles converge towards $j/N = 3/5$.

We now present histogram regression results from three experiments with the EMPL. We start with a toy example intended to build some intuition for the problem at hand, also showing that predicting average histograms with standard loss functions is insufficient when the data exhibits high stochasticity. Then, we consider an application to sports and lastly, we study a computer vision task in γ-ray astronomy. An additional example that considers a bimodal distribution within each bin is provided in the Supplementary Material.

3.1. Toy example: drawing balls from an urn

First, we illustrate our method by means of a toy example, for which the analytic solution can be computed. Minimizing the pinball loss is equivalent to maximizing the likelihood of an asymmetric Laplace distribution (Yu & Moyeed, 2001), which requires the outcome in each bin to be continuous. Whilst interpolation techniques such as *jittering* could be applied to the outcome in the discrete case (and are needed in fact to obtain analytical convergence results; Machado & Santos Silva 2005; Padellini & Rue 2018), we will show in this example that even in the extreme case where the set of possible outcomes consists of the two integers $\{0, 1\}$, the predictions for the quantiles *in practice* behave as expected.

The scenario is the following: we randomly draw x times with replacement from an urn that contains numbered but otherwise identical balls $\mathcal{N} = \{1, \ldots, N\}$ such that each ball has equal probability of being drawn. We keep track of the drawn numbers by adding a tally mark in the respective field (or bin) of a table after each draw before putting the ball back into the urn again. This yields a frequency histogram $(m_j^c)_{j=1}^N$, with m_j^c denoting the number of times the ball j has been drawn. The task of the NN will be to estimate the distribution of the *relative* counts in each bin $(m_j)_{j=1}^N$, where $m_j = m_j^c/x$, depending on the number of draws x.

The total number of counts m_j^c in each bin $j \in \mathcal{N}$ follows a binomial distribution $B(x, p)$, where $p = 1/N$. Since the mean and variance of the relative counts in each bin are given by $\mathbb{E}[m_j] = p$ and $\mathrm{Var}(m_j) = p(1-p)/x$, the relative counts $m_j \to p$ as $x \to \infty$, for all $j \in \mathcal{N}$, implying that each ball will be drawn equally often in the (hypothetical) limit of infinitely many draws. However, for small values of x, the variability of the resulting histograms is high, and in the case of a single draw $x = 1$, $m_j = 0$ in $N - 1$ bins, while $m_j = 1$ in the bin for the drawn number j.

Since the EMPL compares the *cumulative* histograms, we also define the cumulative histogram as $(M_j^c)_{j=1}^N$, where $M_j^c = \sum_{r=1}^j m_r^c$, and similarly for the *relative* cumulative histogram $(M_j)_{j=1}^N$. We write $Y \sim \mathcal{U}\{1, N\}$ for the random variable Y describing a single draw from the urn (i.e. from a discrete uniform distribution between 1 and N). We can determine the CDF for the cumulative counts M_j^c in each bin $j \in \mathcal{N}$ by computing the conditional probability for drawing at most $l \in \{0, \ldots, x\}$ times a number less than or equal $y \in [1, N]$, given by

$$P(\#(Y \le y) \le l \,|\, x) = P(M_j^c \le l \,|\, x)$$
$$= \sum_{m=0}^l \binom{x}{m} p_{\le j}^m (1 - p_{\le j})^{x-m}, \quad (13)$$

where $j = \lfloor y \rfloor$ (only integers can be drawn), $p_{\le j} = j \,/\, N$ is the probability for drawing a number less than or equal y

in a *single* draw, and the probabilities for drawing *exactly* $0, \ldots, l$ times a number $Y \leq y$ need to be summed up to obtain the probability for drawing *at most* l times such a number. Inverting this relation yields the quantiles for the distribution of the value M_j^c (and equivalently M_j) in each bin, conditional on x.

For our numerical experiment, we choose $N = 5$ balls and take the number of draws x itself to be a random variable X, given by $X = \mathrm{round}(\hat{X})$ with $\log_{10}(\hat{X}) \sim U(0, 3)$. We train a simple multilayer perceptron (MLP) containing 2 hidden layers with 128 neurons each, ReLU activation and batch normalization for the hidden layers, and a softmax activation for the output layer to obtain $\tilde{Q}^\theta(X; \tau)$ as described in Sec. 2.5. The NN training consists of 10,000 batch iterations at a batch size of 2,048. We minimize the EMPL with randomly drawn quantile levels τ using an Adam optimizer (Kingma & Ba, 2014). The 2-dimensional inputs to the NN are given by x and τ, and the corresponding 5-dimensional labels $(m_j)_{j=1}^N \,|\, X = x$ are generated by randomly drawing x times from a discrete uniform distribution with range $\mathcal{N}$ and normalizing the resulting histogram. Equivalently (and faster), one can draw from a multinomial distribution with x trials and uniform probability $p_j = p = 1/N$ for all $j \in \mathcal{N}$.

Fig. 2 shows the quantiles of the estimated cumulative relative counts in each histogram bin for $x = 1, 10, 100$, and $1,000$ draws (left panels). The colored lines correspond to the predicted quantiles as indicated in the top panel next to the lines, and the dark gray delimiters show the analytic values for the two most extreme considered quantiles ($\tau = 1/10$ and $9/10$). The right panels depict the true CDFs of the relative cumulative counts in the central bin for each x, i.e. $F(M_3 \,|\, x)$, together with the analytic (dashed) and predicted (solid) quantiles, given by the pre-image of τ (see the horizontal dashed lines) under the CDF.

For a single draw from the urn, i.e. $x = 1$, the only possible values of the histograms are 0 and 1. The cumulative histogram in bin j for a quantile level τ should be 1 if τ is greater than the probability of drawing a number greater than j, i.e. if $\tau > (N - j)/N$, and 0 else. For all values of τ, the MLP correctly determines where the cumulative histogram jumps from 0 to 1. Since the EMD coincides with the EMPL for $\tau = 0.5$, a NN trained by minimizing the EMD predicts $m_j = 1$ for the central bin $j = 3$ and $m_j = 0$ otherwise.[1] In contrast, using the cross-entropy loss for training produces the expected *mean* histogram in each bin ($m_j = 1/N$; independently of the input x), which for $x = 1$ is not representative of any observed histogram with values in $\{0, 1\}$. Therefore, among the considered loss functions, only the EMPL is able to adequately express the full *range* of possible histograms, thanks to the dependence of the NN outputs on the quantile level of interest τ.

As the number of draws x increases, the CDF of M_j gradually becomes narrower (see the right panels for the central bin $j = 3$) and consequently, the quantile ranges converge from both sides towards $p_{\leq j} = j/N$. For all x and τ, the predicted quantiles match their analytic counterparts. Note that in the limit of no stochasticity $x \to \infty$, the NN prediction with the EMPL would become τ-independent and equal its cross-entropy and EMD loss counterparts.

3.2. An application to the football Bundesliga

Week after week, millions of fans around the globe cheer passionately for their favorite football club in the hope of claiming the league title by finishing first at the end of the season. Whilst each club tries its best to win as many matches as possible, their fortune is not entirely in their own hands: 38 points at the end of the Bundesliga (German top-flight division) season 1997−98 were not enough to save Karlsruher SC from being relegated to 2. Bundesliga placed 16th; however, the same number of points would have sufficed for position 13 in season 2001−02, in safe distance from the relegation spots 16−18. Thus, knowledge of the results of a single club *in isolation* is a strong indicator of how well the club fares in terms of the league table, but is not sufficient to determine its position.

We apply the EMPL to the following task: given the list of points that a club has earned in each match during a season $X \in \{0, 1, 3\}^{34}$ (win: 3 points, draw: 1 point, defeat: 0 points; for 34 matches), estimate the histogram $(m_j)_{j=1}^N$ that results from the positions of the club in the league table after each week. Narrow histograms (steep cumulative histograms) indicate few change in the position over the course of the season, while a wide histogram (a gently increasing cumulative histogram) suggests a turbulent season for the respective club in terms of its place in the table. For instance, if a club managed to lead the table throughout the season over 34 weeks, this would result in a histogram with $m_1 = 1$ and $m_j = 0$ for $j = 2, \ldots, 18$ (where 18 is the number of competing clubs).

We use data from all the Bundesliga seasons between 1995−96 (when the 3-points-for-a-win rule was introduced) and 2017−18, keeping the seasons 1998−99, 2006−07, and 2014−15 for testing, while using the other 20 seasons as training data.[2] In order to increase the amount of training data, we "re-play" each training season 1,000 times, randomly permuting the 34 weeks (each of the 18 clubs plays $2 \times 17 = 34$ matches in a season, namely a home and an away match against every other club). Clearly, these artificial seasons converge to the same league table by the end of the season, but the histograms of the table positions after each week are distinct (as is the ordering of the input lists X containing the points from each match).

---

[1] See the Supplementary Material for the expected EMD for $x = 1$.

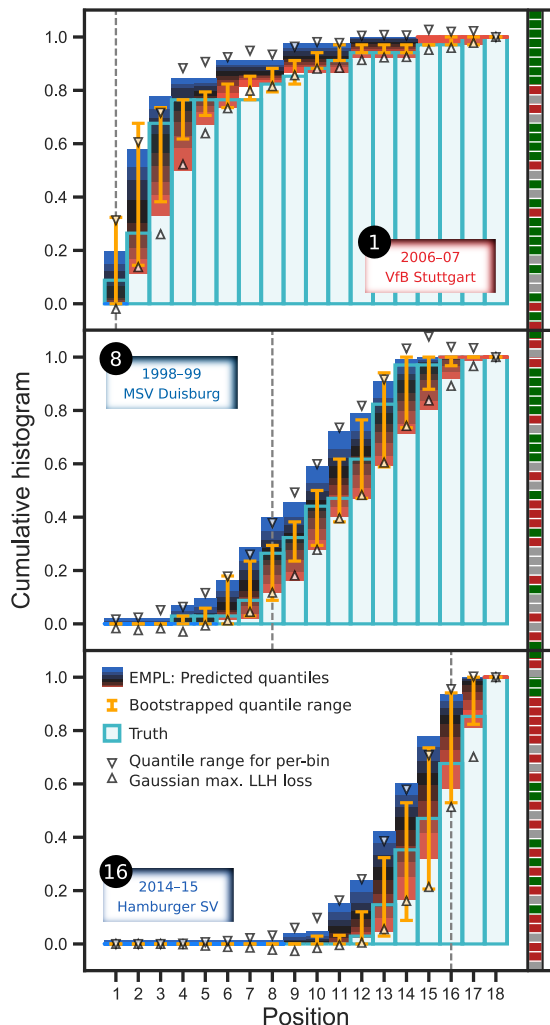[2] Data: www.kaggle.com/thefc17/bundesliga-results-19932018.

Figure 3. Cumulative histograms depicting the distribution of the table position after each week for three clubs (one from each test season). The light-blue histograms represent the truth, and the colored regions show the estimated quantiles from $10 - 90\%$ in steps of 10% with the EMPL. The orange error bars show bootstrapping estimates of the $10 - 90\%$ quantile range. When simply training a NN to estimate bin-wise means and standard deviations for the cumulative histogram assuming a Gaussian likelihood (independently in each bin), the resulting quantile range is not confined to $[0, 1]$ due to the infinite support of the Gaussian distribution and may be non-monotonic (white triangles), which are undesirable properties. The vertical dashed lines and white numbers indicate the position of the club at the end of the respective season (which cannot be inferred from the histograms). The results of all the club's matches during the season (i.e., the input X for the MLP) are illustrated to the right of the histograms (see main text).

We train again a simple MLP with 2 hidden layers with 128 neurons followed by ReLUs and 50% dropouts (Hinton et al., 2012; Srivastava et al., 2014) to prevent overfitting. A softmax activation function produces the relative histograms, consisting of 18 bins corresponding to the league table posi-

tions. We minimize the Smoothed EMPL with $\alpha = 0.005$ over 250 epochs using a batch size of 2,048.

Fig. 3 shows the predictions for the relative cumulative histograms for three clubs (one from each test season). The light-blue histogram corresponds to the true cumulative histogram for the club in the respective season, and the gray dashed line shows the final standing of the club (which cannot be deduced from the histogram). The colored regions indicate the estimated quantile ranges, from 10 to 90% in steps of 10%. The orange error bars show bootstrapping estimates of the bin-wise quantiles obtained by computing hypothetical histograms that would arise had the club obtained the same points in each match in *another* (artificially generated) season. Specifically, we randomly select 200 seasons from our augmented dataset, remove the club whose final number of points is closest (in order to minimize the bias due to situations such as having two champions from different seasons compete against each other, which would bias the histograms towards lower positions), and calculate the bin-wise quantiles over the resulting histograms. The lists of points in each match (i.e. the inputs X fed to the MLP) are illustrated on the right-hand side next to the histograms (1st match at the bottom, 34th match on top; green: 3/win, gray: 1/draw, red: 0/defeat). Although the estimated quantile ranges do not perfectly match their bootstrapping counterparts, their magnitudes are generally similar, and the EMPL enables the quantification of the uncertainty in the distribution of a club's table position over the season, based on other seasons and without any knowledge about the results of the other clubs. Thus, the available domain knowledge that is learned during the training in combination with a limited number of observations allows one to derive a narrow posterior distribution that expresses which histograms are compatible with the observations.

A clear advantage of quantile-based approaches is that they do not assume a specific underlying distribution. We illustrate this by comparing our method with a naive likelihood-based approach for quantifying the bin-wise uncertainty in the histograms, namely a NN trained by simply maximizing the Gaussian log-likelihood for the cumulative histogram $(M_j)_{j=1}^N$ with independent means μ_j and standard deviations σ_j for each bin (resulting in $2 \times 18 = 36$ outputs). While the monotonicity of the mean estimates $(\tilde{\mu}_j)_{j=1}^N$ and the normalization $\tilde{\mu}_N = 1$ are enforced by computing $\tilde{\mu}$ as the cumulative sum of softmax-activated logits (see Sec. 2.5), other quantiles of the cumulative histograms do not need to be monotonic (caused by $\tilde{\sigma}_j \neq \tilde{\sigma}_{j-1}$); also, the values are unbounded due to the infinite support of the Gaussian distribution. In contrast, the EMPL predictions, which do not presume a particular distribution of the histogram values within each bin, lie in $[0, 1]$ and increase monotonically for all τ *by construction*, which are essential properties.
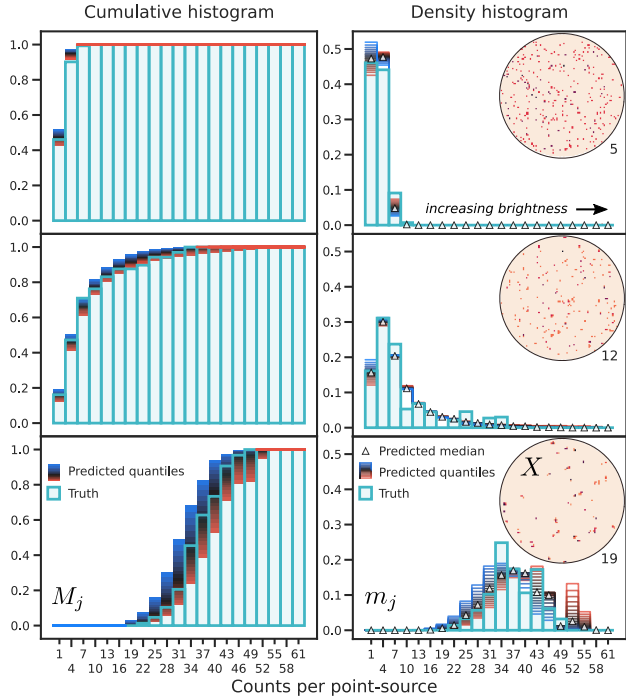
Figure 4. Predicted and true cumulative (left) and resulting density (right) point-source brightness histograms. Projections of the corresponding input photon-count maps X are shown in the inset plots (log-scaled), and the maximum number of counts per pixel (which is considerably lower than the maximum number of counts *per point-source* because of the point spread function), which defines the upper limit of the color map, is indicated next to the maps.

3.3. An example from astrophysics: predicting point-source brightness distributions

Now, we consider a computer vision problem in the field of γ-ray astronomy, namely the estimation of the point-source brightness distribution $(m_j)_{j=1}^N$ given a photon-count map $X \in \mathbb{N}^{n_{\mathrm{pix}}}$ as an input, where n_{pix} is the number of pixels in the region of interest (ROI, taken to be a circle of radius $20°$ here). Specifically, each observed map X_s contains emission from $T_s \in \mathbb{N}$ point-sources, each of which contributes $C_s^t \in \mathbb{N}$ photon counts to the map (for $t = 1, \ldots, T_s$), such that the total number of counts in map s is $C_s^{\mathrm{tot}} = \sum_{t=1}^{T_s} C_s^t$. Binning the counts in the map according to $(C_s^t)_{t=1}^{T_s}$ results in a histogram that characterizes the brightness distribution of the generating point-source population: for each $t = 1, \ldots, T_s$, C_s^t counts are added to the associated bin (for example, for a source t responsible for $C_s^t = 4$ counts, these 4 counts are added to the "$3-5$ counts" bin), implying that the counts from dim (bright) point-sources go to low (high) bins. Once all the counts are distributed, the histogram is normalized to sum up to unity. The task of the NN is to estimate this underlying point-source brightness distribution $(m_j)_{j=1}^N$ from a photon-count map X.

Training loss	MAE[1] ↓	MSE[1] ↓	EM1[2] ↓	EM2[2] ↓	IS [%] ↑
MAE	4.1	8.6	10.8	4.6	95.7
MSE	4.0	**8.1**	10.3	4.2	95.8
XE	4.2	8.6	9.4	4.2	95.6
EMP (**all** τ)	4.0	8.3	8.8	**3.9**	95.8
SEMP (**all** τ)	4.0	8.2	9.0	**3.9**	95.8
EM1 ($\tau = 0.5$)	**3.9**	8.2	**8.7**	**3.9**	**95.9**

[1]: $\times 1{,}000$, [2]: $\times 100$
MAE / MSE: mean absolute / squared error, XE: cross-entropy, (S)EMP: (Smoothed) EMPL (always *evaluated* for $\tau = 0.5$), $\alpha = 0.001$ when smoothed, EM1/2: absolute/squared EMD, IS: histogram intersection (Swain & Ballard, 1991).

Table 1. Different metrics (columns) when evaluating the NN on 512 test maps, for NNs trained using different loss functions (rows). The EM-based losses perform similarly to the per-bin losses in terms of per-bin metrics (MAE/MSE/IS), and achieve better results as measured by the EMD. Training the NN for all quantile levels τ simultaneously barely affects the median accuracy as compared to the EM1 loss ($\tau = 0.5$ only), while yielding much more expressive outputs through *arbitrary* quantiles, thus providing *uncertainties*.

A particularly interesting application is the analysis of the photon-count map from the *Fermi* space telescope (Abdollahi et al., 2020), which contains unexplained excess emission from the center of our Milky Way galaxy (Goodenough & Hooper, 2009) that could possibly be explained by annihilation of dark matter particles (Hooper & Linden 2011). Recently, machine learning methods have opened up a new avenue for the analysis of this excess (Caron et al., 2018; List et al., 2020; Mishra-Sharma & Cranmer, 2020). Whilst an exhaustive study of the *Fermi* map is beyond the scope of this work, we demonstrate that our method is able to estimate the histogram describing the brightness distribution of point-sources in a simple scenario with simulated photon-count maps. We generate 312,500 photon-count maps with the tool NPTFit-Sim (Rodd & Toomey), modeling emission from isotropically distributed point-sources. The photon counts from each source are smeared out by the *Fermi* point spread function over multiple pixels, and each pixel may contain counts from more than one source, making the problem probabilistic and non-trivial. We subsequently discard the maps with less than 1,000 counts and those that contain very bright point-sources with > 60 counts; then, we put aside 1/15th of the remaining maps for testing and use the others as training data.

The input maps X_s are discretized using the HEALPix tessellation of the sphere (Gorski et al., 2005), and we use a resolution set by the parameter $N_{\mathrm{side}} = 256$ (giving $n_{\mathrm{pix}} = 65{,}536$ in our ROI). As proposed by List et al. (2020), we employ a graph-convolutional NN built on the DeepSphere framework (Perraudin et al., 2019; Defferrard et al., 2020), in which the HEALPix sphere is described by a weighted

undirected graph, and the convolution operation is defined by means of the graph Laplacian operator. Our NN is composed of 8 graph-convolutional layers, each followed by maximum pooling, batch normalization, and a ReLU activation, and three fully-connected (FC) layers. The quantile level τ is appended before the first FC layer.

Fig. 4 shows three examples from the testing dataset, for a dim (top), moderate (middle), and bright (bottom) point-source population. The cumulative and density histograms are depicted in light blue (truth) and by colored regions / lines (NN), corresponding to $5 - 95\%$ quantiles in steps of 5%. The white triangles in the right panels are located at the predicted medians. The NN has learned to faithfully recover the underlying brightness histograms. The uncertainties become larger with increasing brightness for equally spaced bins as considered here; however, we found in our experiments that this trend generally reverses when using logarithmically spaced bins.

Table 1 lists several metrics when using different loss functions for the NN training, evaluated on 512 testing maps. We emphasize that in the case of the median $\tau = 0.5$, for which we report our results with the EMPL, the EMPL by construction is *exactly identical* to the EMD (see Eq. (10), up to $1/2$), as the EMPL naturally extends the EMD to arbitrary quantiles. Therefore, one should not expect a higher accuracy when evaluating the EMPL-trained NN for the specific value $\tau = 0.5$ as compared to the EMD-trained NN. In turn, Table 1 shows that training the same NN to estimate *all* the quantiles using the EMPL rather than only the median *barely affects* the accuracy of the median predictions. Thus, the EMPL provides much more expressive outputs that quantify the uncertainties "for free". The EM-based losses (EM1 & EMPL) outperform the bin-wise losses w.r.t. the cross-bin metrics EM1 and EM2, while performing similarly in terms of the bin-wise metrics.

Whilst labeled histogram data may be difficult to acquire or might not be available at all in some applications, this astrophysical example belongs to the important class of problems where labeled training data can be obtained (for instance using a simulator), but *recovering* the underlying histogram from real data is a challenging task that can be tackled by Deep Learning methods. For instance, CNNs are able to assess the real photon-count map of the sky on multiple scales, which can potentially give rise to more robust results in the presence of mismodeling on large angular scales (List et al., 2020), whereas statistical methods typically rely on an approximation of the likelihood that treats each pixel independently (e.g. Mishra-Sharma et al. 2017).

4. Conclusions

We have presented a method for the NN-based regression of histograms from input images or other data. Our approach is based on minimizing a novel loss function, which we call the *Earth Mover's Pinball Loss* (EMPL), rooted in transportation theory as well as in quantile regression. This loss function is an asymmetric generalization of the EMD that allows for the regression of *arbitrary quantiles* of the cumulative histogram in each bin, harnessing the idea of the pinball loss. In the particular case of the median ($\tau = 0.5$), our loss function reduces to the EMD. We have demonstrated the effectiveness of our method in a toy example, a football-related task, and a problem in γ-ray astronomy. The accuracy of the estimated median histogram is very similar to the standard EMD case, and the prediction of arbitrary other quantiles comes at almost no additional cost (the increase in walltime for training is $< 10\%$). Given the vast range of applications where histograms are used, there is a great potential for Deep Learning methods to provide accurate, fast, and reliable histogram predictions. The EMPL is easy to implement (see the Supplementary Material), and we expect it to be particularly useful for tasks where the entire *distribution* of possible histograms is of interest such as rain forecasts ("what's the probability that it rains more than 10 mm tomorrow?") or radiation treatment planning ("how certain can we be that 20% of the cancerous organ should receive a dose of 30 Gy?"). Possible extensions of our work include multidimensional histograms, incorporating epistemic uncertainties (e.g. Gal & Ghahramani 2016), flexible ground distances, and the application to parameterized continuous (i.e. unbinned) distributions.

---

[3]https://github.com/cloudpipe/cloudpickle
[4]https://github.com/holoviz/colorcet   [5]https://inkscape.org/

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *preprint (arXiv:1603.04467)*, 2016.

Abdollahi, S., Acero, F., Ackermann, M., Ajello, M., et al. Fermi Large Area Telescope Fourth Source Catalog. *The Astrophysical Journal Supplement Series*, 247:33, 2020.

Aigner, D. J., Amemiya, T., and Poirier, D. J. On the Estimation of Production Frontiers: Maximum Likelihood Estimation of the Parameters of a Discontinuous Density Function. *International Economic Review*, 17:377–396, 1976.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *ICML*, pp. 214–223, 2017.

Bellerby, T. J. Satellite rainfall uncertainty estimation using an artificial neural network. *Journal of Hydrometeorology*, 8:1397–1412, 2007.

Caron, S., Gómez-Vargas, G. A., Hendriks, L., and de Austri, R. R. Analyzing γ rays of the Galactic Center with deep learning. *Journal of Cosmology and Astroparticle Physics*, 2018:058, 2018.

Chan, T. H., Jia, K., Gao, S., Lu, J., et al. PCANet: A Simple Deep Learning Baseline for Image Classification? *IEEE Transactions on Image Processing*, 24:5017–5032, 2015.

Chollet, F. et al. Keras. https://keras.io, 2015.

Defferrard, M., Milani, M., Gusset, F., and Perraudin, N. DeepSphere: a graph-based spherical CNN. *ICLR*, 2020.

Ferguson, T. S. *Mathematical statistics: A decision theoretic approach*, volume 1. Academic press, 2014.

Fox, M. and Rubin, H. Admissibility of Quantile Estimates of a Single Location Parameter. *The Annals of Mathematical Statistics*, 35:1019–1030, 1964.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *ICML*, pp. 1651–1660, 2016.

Geng, X. Label Distribution Learning. *IEEE Transactions on Knowledge and Data Engineering*, 28:1734–1748, 2016.

Goodenough, L. and Hooper, D. Possible Evidence for Dark Matter Annihilation in the Inner Milky Way from the Fermi Gamma Ray Space Telescope. *preprint (arXiv:0910.2998)*, 2009.

Gorski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., et al. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *The Astrophysical Journal*, 622:759, 2005.

Hatalis, K., Lamadrid, A. J., Scheinberg, K., and Kishore, S. A Novel Smoothed Loss and Penalty Function for Noncrossing Composite Quantile Estimation via Deep Neural Networks. *preprint (arXiv:1909.12122)*, 2019.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., et al. Improving neural networks by preventing co-adaptation of feature detectors. *preprint (arXiv:1207.0580)*, 2012.

Hooper, D. and Linden, T. Origin of the gamma rays from the Galactic Center. *Physical Review D*, 84:123005, 2011.

Hou, L., Yu, C.-P., and Samaras, D. Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks. *preprint (arXiv:1611.05916)*, 2016.

Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9:90–95, 2007.

Jin, X., Wu, L., Li, X., Chen, S., et al. Predicting aesthetic score distribution through cumulative Jensen-Shannon divergence. *AAAI*, pp. 77–84, 2018.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *preprint (arXiv:1412.6980)*, 2014.

Koenker, R. and Bassett, G. Regression Quantiles. *Econometrica*, 46:33–50, 1978.

Koenker, R. and Hallock, K. F. Quantile Regression. *Journal of Economic Perspectives*, 15:143–156, 2001.

Lam, S. K., Pitrou, A., and Seibert, S. Numba: A LLVM-Based Python JIT Compiler. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015.

Levina, E. and Bickel, P. The Earth Mover's distance is the Mallows distance: Some insights from statistics. *ICCV*, 2:251–256, 2001.

List, F., Rodd, N. L., Lewis, G. F., and Bhat, I. Galactic Center Excess in a New Light: Disentangling the γ-Ray Sky with Bayesian Graph Convolutional Neural Networks. *Physical Review Letters*, 125:241102, 2020.

Liu, Z., Chen, X., Men, K., Yi, J., et al. A deep learning model to predict dose–volume histograms of organs at risk in radiotherapy treatment plans. *Medical Physics*, pp. 5467–5481, 2020.

Machado, J. A. and Santos Silva, J. M. Quantiles for counts. *Journal of the American Statistical Association*, 100:1226–1237, 2005.

McKerns, M. M., Strand, L., Sullivan, T., Fang, A., and Aivazis, M. A. G. Building a framework for predictive science. *Proceedings of the 10th Python in Science Conference*, 2011.

Mishra-Sharma, S. and Cranmer, K. Semi-parametric γ-ray modeling with Gaussian processes and variational inference. *preprint (arXiv:2010.10450)*, 2020.

Mishra-Sharma, S., Rodd, N. L., and Safdi, B. R. NPTFit: A Code Package for Non-Poissonian Template Fitting. *The Astronomical Journal*, 153:253, 2017.

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Paul, W., Jordan, M. I., and Stoica, I. Ray: A distributed framework for emerging AI applications. *CoRR*, 2017.

Newey, W. K. and Powell, J. L. Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55:819–847, 1987.

Nguyen, H. V. and Vreeken, J. Non-parametric Jensen-Shannon divergence. *Machine Learning and Knowledge Discovery in Databases*, 9285:173–189, 2015.

Nicholls, N., Drosdowsky, W., and Lavery, B. Australian rainfall variability and change. *Weather*, 52:66–72, 1997.

Novak, C. L. and Shafer, S. A. Anatomy of a color histogram. *CVPR*, pp. 599–605, 1992.

Oliphant, T. E. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

Oresme, N. *Tractatus de latitudinibus formarum*. 1486.

Padellini, T. and Rue, H. Model-aware Quantile Regression for Discrete Data. *preprint (arXiv:1804.03714)*, 2018.

Pearson, K. X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. (A.)*, 186:343–414, 1895.

Peeples, J., Xu, W., and Zare, A. Histogram Layers for Texture Analysis. *preprint (arXiv:2001.00215)*, 2020.

Perraudin, N., Srivastava, A., Lucchi, A., Kacprzak, T., et al. Cosmological N-body simulations: a challenge for scalable generative models. *Computational Astrophysics and Cosmology*, 6:5, 2019.

Ramdas, A., Trillos, N., and Cuturi, M. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19:47, 2017.

Rebetez, J., Satizábal, H. F., Mota, M., Noll, D., et al. Augmenting a convolutional neural network with local histograms - A case study in crop classification from high-resolution UAV imagery. *ESANN 2016*, pp. 515–520, 2016.

Rodd, N. and Toomey, M. NPTFit-Sim. URL https://github.com/nickrodd/NPTFit-Sim.

Rubner, Y., Tomasi, C., and Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.

Saadl, H., Ismaie, A. P., Othmanl, N., Jusohl, M. H., et al. Recognizing the ripeness of bananas using artificial neural network based on histogram approach. *ICSIPA09*, pp. 536–541, 2009.

Sedighi, V. and Fridrich, J. Histogram layer, moving convolutional neural networks towards feature-based steganalysis. *Electronic Imaging*, pp. 50–55, 2017.

Sharma, K., Gold, M., Zurbruegg, C., Leal-Taixe, L., and Wegner, J. D. HistoNet: Predicting size histograms of object instances. *WACV*, pp. 3637–3645, 2020.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR*, 15:1929–1958, 2014.

Swain, M. J. and Ballard, D. H. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991.

Tagasovska, N. and Lopez-Paz, D. Single-Model Uncertainties for Deep Learning. *NeurIPS*, pp. 6417–6428, 2019.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

Ustinova, E. and Lempitsky, V. Learning deep embeddings with histogram loss. *NIPS*, pp. 4170–4178, 2016.

Villani, C. *Topics in optimal transportation*. American Mathematical Soc., 2003.

Villani, C. *Optimal Transport*, volume 338. Springer, Berlin, Heidelberg, 2009.

Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272, 2020.

Wang, Z., Li, H., Ouyang, W., and Wang, X. Learnable Histogram: Statistical Context Features for Deep Neural Networks. *ECCV*, pp. 246–262, 2016.

Waskom, M. et al. mwaskom/seaborn: v0.8.1 (10.5281/zen-odo.883859), 2017.

Weeks, J. R. *Population: An introduction to concepts and issues*. Cengage Learning, 2020.

Yu, K. and Moyeed, R. A. Bayesian quantile regression. *Statistics and Probability Letters*, 54:437–447, 2001.

Zheng, S. Gradient descent algorithms for quantile regression with smooth approximation. *IJMLC*, 2:191–207, 2011.

Zholus, A. and Putin, E. Continuous Histogram Loss: Beyond Neural Similarity. *preprint (arXiv:2004.02830)*, 2020.

Zonca, A., Singer, L., Lenz, D., Reinecke, M., Rosset, C., Hivon, E., and Gorski, K. healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in Python. *Journal of Open Source Software*, 4:1298, 2019.