Bayesian Active Learning For Censored Regression

Frederik Boe Hüttel, Christoffer Riis,
Filipe Rodrigues (✉), and Francisco Pereira

Technical University of Denmark rodr@dtu.dk

Abstract. Bayesian active learning is based on information theoretical approaches that focus on maximising the information that new observations provide to the model parameters. This is commonly done by maximizing the Bayesian Active Learning by Disagreement (BALD) acquisition function. However, it is challenging to estimate BALD when the new data points are subject to censorship, where only clipped values of the targets are observed. To address this, we derive the entropy and the mutual information for right-censored distributions and derive the BALD objective for active learning in censored regression ($\mathcal{C}$-BALD). We propose a novel modeling approach to estimate the $\mathcal{C}$-BALD objective and use it for active learning in the censored setting. Across a wide range of datasets and models, we demonstrate that $\mathcal{C}$-BALD outperforms other Bayesian active learning methods in censored regression.

1 Introduction

Active learning is a framework where a model learns from a small amount of labeled data and chooses the data it wants to acquire a label for [43]. This acquisition of new data points is done iteratively to improve the model's predictive performance and reduce model uncertainty [33]. This naturally poses the challenge: which new data points can improve the model the most? Information theoretical approaches are often the basis to solve this challenge by reasoning about the information that new labels can provide to the model's parameters [34]. A widely used strategy in active learning is the Bayesian Active Learning by Disagreement (BALD) acquisition function, which identifies new data points by estimating the mutual information between the model parameters and the acquired labels [17]. BALD has demonstrated effectiveness across various domains such as computer vision [8], natural language processing [45], and survival analysis [37]. However, applying BALD to censored regression tasks introduces unique challenges, where labels are only partially observed due to censoring.

Censored data arises in scenarios where certain observations are incomplete or "clipped" due to limitations in the measurement process [4]. This phenomenon has substantial implications across multiple critical real-world domains. A particularly important context where censoring plays a key role is the ubiquitous predict-then-optimize framework, where systematic bias introduced by censorship can lead to suboptimal decision-making. Let us illustrate this problem

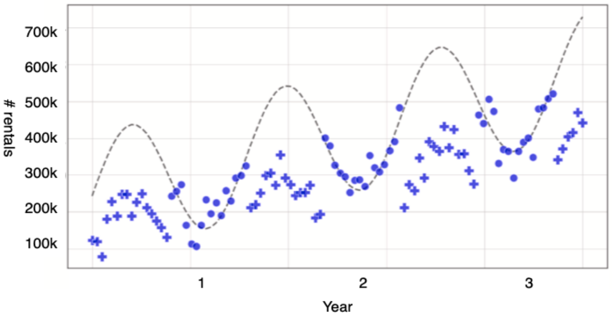

Fig. 1. Evolving shared bikes demand over a 3-year period. The dashed black line represents the (unobserved) true demand. The blue circles denote non-censored observations and the blue crosses represent censored observations (due to limited supply). In the context of bike-sharing network expansion, obtaining new observations of demand in an arbitrary location is expensive since it may involve buying new equipment.

through an application from the transportation domain. Consider the synthetic dataset shown in Figure 1 mimicking bike sharing rentals (demand) for an expanding network of docking stations through time, where the blue markers represent the number of observed rentals (censored observations), and the dashed line represents the unobserved “true demand". Every year, the number of users increases in the population, albeit with seasonal fluctuations. However, the available bicycles (supply) are limited and their number often reaches zero in certain locations, thus resulting in censored observations of the “true demand” as studied, for example, in [20]. In the context of this application, the predict-then-optimize cycle consists of periodically re-estimating the demand and placing new bikes (or docking stations) in areas to minimize the difference between predicted demand and supply. Although costly, additional supply can be placed in areas with high uncertainty to “probe" the demand, i.e., uncover unobserved demand, thus resulting in an active learning setup.

Another example with important real-world implications in the energy domain is electrical vehicle charging infrastructure expansion [11]. The decision of where to place new infrastructure, such as additional fixed or mobile chargers, hinges on these censored observations, aiming to minimize the mismatch between predicted demand and available supply. Obtaining new observations (e.g., via placing new infrastructure) can aid in uncovering the spatio-temporal distribution of the true (uncensored) demand, but it can be extremely costly.

The predict-then-optimize framework is common across various domains with real-world applications, often involving censored data and requiring strategic decision-making under uncertainty. For instance, sensor networks in environmental monitoring, such as oceanography and forestry, face limitations where data collection is constrained by device capabilities. In these cases, models optimize sensor placement in dynamic environments to maximize information gain despite censored/noisy observations. [38] explore Bayesian Gaussian processes for processing sensor data in real-time, accounting for the censoring inherent in environmental or technical constraints. Similarly, [35] examines optimal sensing policies for dynamic decisions based on censored or incomplete data. [12] discuss strategies in cognitive radio systems, where censorship arises from bandwidth limitations or interference, affecting the system's ability to observe the full environment. Emergency response logistics face similar challenges, where resources like medical supplies or personnel must be deployed based on censored demand data. During natural disasters or pandemics, real-time demand for critical resources is often unobserved due to limited availability, creating uncertainty in supply chain optimization. [32] apply Bayesian methods to update demand information, while [49] focus on optimizing emergency logistics in epidemics by accounting for demand urgency.

Common for all these applications is the high cost associated with gathering new data or labels, while maintaining good estimates of the true targets. For instance, in the case of electric vehicle charging networks, obtaining additional observations implies moving or building new infrastructure. Similarly, in online advertisement bidding systems, exploring new strategies comes with a trade-off: the cost of exploration versus the potential gains from exploitation. This balance between acquiring more information and managing costs is crucial across various domains, influencing decision-making under uncertainty and resource constraints. Therefore, this work introduces a novel approach to Bayesian active learning in the context of censored regression. We extend the BALD acquisition function to handle censored observations, formulating the Censored-BALD ($\mathcal{C}$-BALD) acquisition function. This function quantifies the mutual information between censored data points and the model parameters, allowing us to efficiently select the most informative observations, even when they are subject to censoring. By explicitly modeling the censoring process and incorporating it into the active learning framework, we aim to reduce uncertainty in censored regression tasks, providing a principled method to address these complex real-world problems.

2 Background & Setting

We are interested in the supervised learning of a probabilistic regression model, $p(y_i^*|\mathbf{x}_i, \theta)$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ for $d \geq 1$, $y_i^* \in \mathcal{Y}^* \subseteq \mathbb{R}$, and θ is a set of stochastic model parameters. We assume that we can sample a set of model parameters, θ, from the posterior distribution $p(\theta|\mathcal{D})$. We consider the special regression case, where y_i^* is subject to censoring, meaning that for some observations in our dataset, y_i^* is unknown. Specifically, we consider *right-censored* data, which means that instead of observing y^*, we observe $y_i = \min(y_i^*, z_i)$, where $z_i \in \mathcal{Z} \subseteq \mathbb{R}$ is a threshold value of y_i. In addition, we also observe a censoring indicator $\ell_i = \mathbb{1}\{y_i^* \leq z_i\}$, which indicates whether y_i is censored or not. A censored dataset of size n can thus be denoted $\mathcal{D} = \{\mathbf{x}_i, y_i, \ell_i\}_{i=1}^n$.

In the case of censored regression, the objective is to infer the true distribution $p(y_i^*|\mathbf{x}_i, \theta)$ and the model parameters, θ, based on the censored dataset $\mathcal{D}$. In censored regression, one typically assumes that the distributions of $p(y_i^*|\mathbf{x}_i)$ and $p(z_i|\mathbf{x}_i)$ are independent given the covariates, $\mathbf{x}_i$ [46]. This assumption is more general than other assumptions, such as fixed-value censoring, i.e., $z_i = \text{constant}, \forall i$ [40]. We formally state this assumption as follows:

Assumption 1. *(Independent censoring) Conditioned on the covariates,* $\mathbf{x}_i$*, the censoring distribution and the true distribution of the target are independent. That is* $y_i^* \perp z_i|\mathbf{x}_i$.

Under Assumption 1, we obtain the following densities for $p(y^*)$ and $p(y)$,

$$p(y^*|\mathbf{x}, \theta) = \varphi\left(y^*|\mathbf{x}, \theta\right), \tag{1}$$

$$p(y|\ell, \mathbf{x}, \theta) = \varphi\left(y|\mathbf{x}, \theta\right)^{\ell} \left(1 - \Phi\left(y|\mathbf{x}, \theta\right)\right)^{(1-\ell)}, \tag{2}$$

where Φ is the Cumulative Distribution Function (CDF) and φ is the Probability Density Function (PDF) of $p(y^*|\mathbf{x}, \theta)$. Since we do not have access to the non-censored dataset, y^*, we can only estimate θ through their censored counterparts, y. Equation 2, also known as Tobit likelihood [46], thus models the joint distribution of censored ($\ell = 1$) and censored data points ($\ell = 0$). The corresponding log-likelihood loss function for right-censored models simplifies to:

$$\mathcal{L}_{\mathrm{C}}(\theta) = -\sum_{i \in \mathcal{D}} \Big(\underbrace{\ell_i \log\left(\varphi\left(y_i|\mathbf{x}_i, \theta\right)\right)}_{\text{Observed loss}} + \underbrace{(1 - \ell_i) \log\left(1 - \Phi\left(y_i|\mathbf{x}_i, \theta\right)\right)}_{\text{Censored loss}} \Big), \tag{3}$$

While we focus on right-censoring, left-censoring (i.e. $y_i = \max(y_i^*, z_i)$) can be handled by inverting $y_i, \forall i$.

We will assume $p(y_i^*|\mathbf{x}_i, \theta)$ to be Gaussian, such that $p(y_i^*|\mathbf{x}_i, \theta) = \mathcal{N}(\mu_i^*, \sigma_i^{2*}|\mathbf{x}_i, \theta)$. As a consequence, $p(y|\ell, \mathbf{x}_i, \theta)$ will be a mixture model that reduces to a Gaussian when all $\ell = 1$.

Using the loss in Equation 3 we can fit a model of $p(y_i^*|\mathbf{x}_i, \theta)$, as long as its PDF φ and CDF Φ are well-defined. Since our focus is on Bayesian active learning, and concretely BALD, we consider the broad class of Bayesian models. Common choices include deep ensembles [28] and neural networks with stochastic parameters [7,44].

2.1 Active Learning

In the supervised regression setting, active learning involves selecting which labels to acquire during training to increase the model performance [34,43]. It maximizes an acquisition function, which captures the utility of acquiring the label for a given input [26]. We are interested in such settings, but where the data points are subject to censoring. Typically, one starts with a small training dataset, $\mathcal{D}^{\text{train}} = \{(\mathbf{x}_i, y_i, \ell_i)\}_{i=1}^n$, which is used to train a probabilistic model with likelihood $p(y_i^*|\mathbf{x}_i, \theta)$. Then, from a larger (finite or infinite) pool of future unlabelled data, $\mathcal{D}^{\text{pool}} = \{\mathbf{x}_i\}_{t=1}^m$, the model is used to actively select $\mathbf{x}_i$

to acquire a label for [26]. Once the label is acquired, the sample is added to the training set. In the pool, $\mathcal{D}^{\text{pool}}$, the censorship status of new observations is *unknown*, i.e., during acquitions of new observations, both y_i and ℓ_i are unknown [37]. Thus, acquiring new labels involves obtaining its label y_i alongside its censorship status ℓ_i [48].

2.2 Bayesian Experimental design

Bayesian experimental design is a formal framework for quantifying the information gained from an experiment [31]. In active learning, we can view the input $\mathbf{x}_i$ as the design of an experiment and the acquired label y_i as the experiment's outcome and formalize the information gained from observing y_i [2]. Let θ be the quantity we are trying to infer. Given a prior (or the most recent knowledge), $p(\theta)$, and a likelihood function, $p(y_i|\mathbf{x}_i, \theta)$, then we can quantify the information gain (IG) in θ due to an acquisition of $(\mathbf{x}_i, y_i)$, as the reduction in Shannon entropy in θ that results from observing $(\mathbf{x}_i, y_i)$:

$$\mathrm{IG}_\theta(\mathbf{x}_i, y_i) = \mathrm{H}[p(\theta)] - \mathrm{H}[p(\theta|\mathbf{x}_i, y_i)]\,. \tag{4}$$

Since y_i is a random variable, the expected information of y_i can be computed across multiple simulated outcomes using

$$p_\theta(y_i|\mathbf{x}_i) = \mathbb{E}_{p(\theta)}[p(y_i|\mathbf{x}_i, \theta)], \tag{5}$$

which leads to the expected information gain,

$$\mathrm{EIG}_\theta(\mathbf{x}_i) = \mathbb{E}_{p_\theta(y_i|\mathbf{x}_i)}\left[\mathrm{H}\left[p(\theta)\right] - \mathrm{H}\left[p(\theta|\mathbf{x}_i, y_i)\right]\right]\,. \tag{6}$$

This is the expected reduction in uncertainty of θ after conditioning on $(\mathbf{x}_i, y_i)$. Equivalently, it is the mutual information between θ and y_i given $\mathbf{x}_i$, denoted $\mathrm{I}\left[y_i, \theta|\mathbf{x}_i\right]$ [2].

2.3 Bayesian active learning

The expected information gain has often been the basis for Bayesian active learning, seeking to acquire data points that provide high information gain in the model parameters θ. This acquisition function is referred to as the *Bayesian Active Learning by Disagreement* (BALD) [17]:

$$\begin{aligned}\mathrm{BALD}\left(\mathbf{x}_i\right) &= \mathbb{E}_{p_\theta(y_i|\mathbf{x}_i)}[\mathrm{H}[p(\theta)] - \mathrm{H}[p(\theta|\mathbf{x}_i, y_i)]] \\ &= \mathbb{E}_{p(\theta)}[\mathrm{H}[p(y_i|\mathbf{x}_i)] - \mathrm{H}[p(y_i|\mathbf{x}_i, \theta)]] \\ &= \mathrm{H}[p(y_i|\mathbf{x}_i)] - \mathbb{E}_{p(\theta)}[\mathrm{H}[p(y_i|\mathbf{x}_i, \theta)]]\,,\end{aligned} \tag{7}$$

where the unconditional entropy is obtained by marginalizing over the parameters θ,

$$\mathrm{H}[p(y_i|\mathbf{x}_i)] = \mathrm{H}[\mathbb{E}_{p(\theta)}[p(y_i|\mathbf{x}_i, \theta)]]\,. \tag{8}$$

The BALD score is often used when the update to the model parameters is non-Bayesian, for example, when applying Monte Carlo dropout in a neural network [8], which we use to approximate the marginalizing over the parameters θ in Equation 8. For Bayesian active learning without censoring, the BALD acquisition function can be used for classification and regression methods, as the entropies are well-defined for these tasks [8,21].

3 Censoring and Information

Ideally, we would still like to use the BALD objectives for active learning in the censored data case. However, we must consider that, for a new observation $\mathbf{x}_i$ in the pool, the corresponding label y_i can provide a varying amount of information for the distribution of y_i and θ depending on the censorship status of the label [1,14,15,16]. To use the EIG and BALD acquisition functions, we will derive the information (Shannon entropy) for a model trained with Equation 3. Using the derived entropy, we extend the BALD objective to the censored case and use this as an acquisition function for Bayesian active learning in this setting. For the entropy equations in the following, we omit the dependency on $\mathbf{x}_i$, i, and θ for readability.

3.1 Censored information

In the case of non-censorship, the amount of information that y provides to the continuous distribution $p(y)$ corresponds to the Shannon differential entropy, defined as,

$$\mathrm{H}[p(y)] = -\int p(y) \log p(y) dy \ = -\mathbb{E}_{y \sim p(y)} \left[\log p(y) \right]. \tag{9}$$

If we consider the density of a (right) censored distribution introduced in Equation 2, we can formulate the entropy for a censored distribution into the following entropy,

$$\mathrm{H}[p(y|\ell)] = -\mathbb{E}_{y \sim p(y|\ell)} \left[\ell \log \varphi(y) + (1-\ell) \log(1 - \varPhi(y)) \right]. \tag{10}$$

This entropy naturally reflects our assumption of censored observations that the reduction in entropy from observing y is conditioned on its censorship status. Using the fact that if y is censored, then $y = z$ and if y is not censored, $y = y^*$, then we can reformulate the entropy as,

$$\mathrm{H}\left[p(y|\ell)\right] = -\mathbb{E}_{y \sim p(y|\ell)} [\ell \log \varphi\left(y^*\right) + (1-\ell) \log(1 - \varPhi(z))]. \tag{11}$$

However, the censoring information (both z and ℓ) is *unknown* during acquisitions of new data points. To use this entropy for active learning, we propose treating the indicator variable ℓ as a binary random variable. Later, we describe

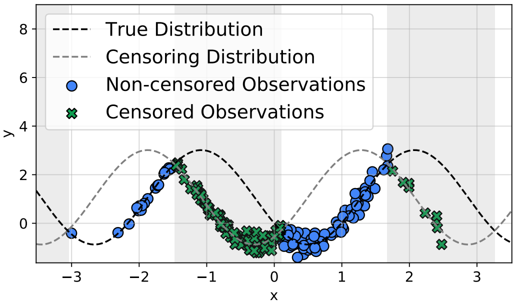

Fig. 2. Illustration of a 1-dimensional censored dataset, in which the dashed black line represents the underlying function that generated the data. The (blue) circles denote non-censored observations, while the (green) crosses represent observations that have been censored. The grey background indicates areas where the observations are censored.

our approach to modeling ℓ. Having a model of $p(\ell)$, we propose to approximate Eq. 11 with its expectation with respect to $p(\ell)$:

$$\begin{aligned}\mathrm{H}[p(y|\ell)] &\approx \mathbb{E}_{p(\ell)}\left[\mathrm{H}[p(y|\ell)]\right] \\ &= -\mathbb{E}_{y\sim p(y|\ell)}[p(\ell)\log\varphi\left(y^{*}\right) + (1-p(\ell))\log(1-\mathit{\Phi}(z))]\,. \end{aligned} \tag{12}$$

3.2 Expected information gain in censored acquisitions

We can use the derived entropy to calculate the information that newly observed targets y_i will provide to the parameters of a Bayesian model. However, the acquisition of new labels not only requires obtaining new values of y_i, but it also involves acquiring new censoring indicators ℓ_i [37]. Consequently, it is necessary to account for the mutual information between y_i and θ and consider the information provided by ℓ_i. As a result, we jointly compute the mutual information between the set (y_i, ℓ_i) and θ. This leads to the following mutual information (derivation in Appendix A.1),

$$\begin{aligned}\mathcal{C}\text{-BALD}(\mathbf{x}_i) &= \mathrm{I}\left[(y_i,\ell_i),\theta|\mathbf{x}_i\right] \\ &= \mathbb{E}_{p(\theta)}[\mathrm{H}[p_\theta(y_i,\ell_i|\mathbf{x}_i)] - \mathrm{H}[p(y_i,\ell_i|\mathbf{x}_i,\theta)]]\,. \\ &= \mathrm{I}\left[y_i,\theta|\ell_i,\mathbf{x}_i\right] + \mathrm{I}\left[\ell_i,\theta|\mathbf{x}_i\right]\ . \end{aligned} \tag{13}$$

Therefore, in the censored regression case, the information gained from observing y_i and ℓ_i is the information provided by observing the label y_i given the censoring indicator ℓ_i, plus the information from observing the censoring indicator ℓ_i. The mutual information criteria can be computed similarly to the BALD objective using the derived entropies,

$$\begin{aligned}\mathrm{I}\left[y_i,\theta|\ell_i,\mathbf{x}_i\right] &= \mathbb{E}_{p(\theta)}[\mathrm{H}[p_\theta(y_i|\ell_i,\mathbf{x}_i)] - \mathrm{H}[p(y_i|\ell_i,\mathbf{x}_i,\theta)]]\,, \\ \mathrm{I}\left[\ell_i,\theta|\mathbf{x}_i\right] &= \mathbb{E}_{p(\theta)}[\mathrm{H}[p_\theta(\ell_i|\mathbf{x}_i)] - \mathrm{H}[p(\ell_i|\mathbf{x}_i,\theta)]]\,. \end{aligned} \tag{14}$$

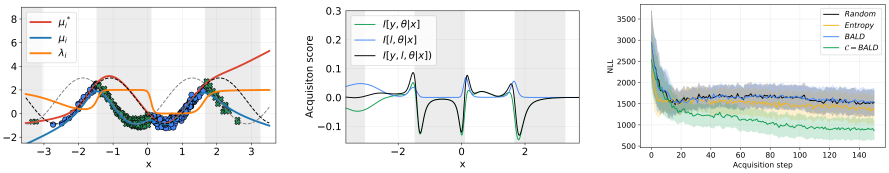

Fig. 3. Left): Overview of the fit of the proposed modeling approach on the 1-D synthetic dataset. Red: Estimated distribution of the true function, $p(y_i^*|\mathbf{x}_i, \theta)$. Blue: Estimated distribution of the observed values, $p(y_i|\mathbf{x}_i, \theta)$. Orange: Estimated probability of being censored, $p_\theta(\ell_i|\mathbf{x}_i)$ (scaled between 0 and 2 for illustration purposes). **Middle):** The mutual information calculations for the label y and the censoring status. Grey areas indicate areas with complete censoring. Most information comes from the cross-over point between the censored and non-censored values. **Right):** The right censored NLL for the models across different acquisition functions on the synthetic dataset (mean ± standard error). $\mathcal{C}$-BALD achieves the best overall fit on the test set.

4 Information Gain in $\mathcal{C}$-BALD

A fundamental challenge of $\mathcal{C}$-BALD for active learning is that the censored regression model, estimated using the censored loss function from Equation 3, only approximates the parameters for the distribution of y_i^*, and its corresponding PDF φ and CDF Φ. This means that during acquisition, there is no knowledge of the potential censoring status of new observations ℓ_i, which is required to compute the mutual information (Equation 13), and there is no knowledge of the potential censoring threshold z_i, which is required to compute the entropy of y_i (Equation 12). Therefore, applying $\mathcal{C}$-BALD in practice is not straightforward. To overcome these challenges, we propose explicitly modeling the probability of being censored ℓ_i and the censoring threshold z_i as described below.

Modelling of ℓ_i: Recall that the censoring indicator $\ell_i = \mathbb{1}\{y_i^* \leq z_i\}$ is observed for each data point in a censored dataset. It is a binary indicator of whether the observations are censored or not. Using a neural network, we propose to fit the distribution of $p(\ell_i|\mathbf{x}_i, \theta)$. Concretely, we parameterise $p(\ell_i|\mathbf{x}_i, \theta)$ as a Bernoulli distribution $\text{Ber}(\lambda_i|\mathbf{x}_i, \theta)$, and infer the parameters θ using the binary cross entropy loss ($\mathcal{L}_{\text{BCE}}(\theta)$). Consequently, this explicit modelling of ℓ_i allows us to approximate the mutual information $\text{I}[\ell_i, \theta|\mathbf{x}_i]$ required for Equation 13 ($\mathcal{C}$-BALD).

Modelling of z_i and y_i: Explicit modeling of z_i is more challenging, as it is not fully observed (similarly to y_i^*). However, notice that for computing the conditional entropy in Equation 12, we are only interested in the value of z_i for the case when $\mathbf{x}_i$ is subject to censoring, i.e. $\ell_i = 0$, in which case $y_i = z_i$. This implies that we directly observe the true values of z_i when we have censored data points. Therefore, we propose to also explicitly model y_i using a standard Gaussian distribution, $\tilde{p}(y_i|\mathbf{x}_i, \theta) = \mathcal{N}(\mu_i, \sigma_i^2|\mathbf{x}_i, \theta)$, estimated with the maximum likelihood (with loss $\mathcal{L}_{\mathrm{GAUSS}}(\theta)$)[1]. Notice that $\tilde{p}(y_i|\mathbf{x}_i, \theta)$ represents the distribution of the observations regardless of censorship, while $p(y_i^*|\mathbf{x}_i, \theta)$ represents that actual latent distribution we are trying to model. By explicitly modelling both $\tilde{p}(y_i|\mathbf{x}_i, \theta_i)$ and $p(\ell_i|\mathbf{x}_i, \theta)$, for censored observations (when $\ell_i = 0$), $\mathbb{E}[y_i] = \mu_i$ provides an estimate of z_i, while for uncensored observations $\mathbb{E}[y_i^*] = \mu^*$. This approach thus allows us to estimate z_i for censored cases, which is crucial for the entropy calculations described in Equation 12. An example of this can be seen in Figure 3, where $\tilde{p}(y_i|\mathbf{x}_i, \theta_i)$ follows the data points, also for censored observations, thereby estimating z_i.

Entropy estimation: With this explicit modeling approach, we can approximate information that new observations provide to the parameters of the censored model.

$$\mathrm{H}[p(y_i)] \approx -\,\mathbb{E}_{y_i\sim p(y_i)}[p_\theta(\ell_i|\mathbf{x}_i)\log\varphi\,(y_i) + (1 - p_\theta(\ell_i|\mathbf{x}_i))\log(1 - \Phi(\mu_i))]\,. \tag{15}$$

4.1 Summary and implementation details

We want to use the mutual information between observations of (y_i, ℓ_i) and the model parameters θ to acquire new labels to reduce model uncertainty about y_i^*. Since the distribution of y^* is not fully observed, we use the entropy defined in Equation 12 to compute the mutual information. However, the entropy relies on the knowledge of unknown variables z_i and ℓ_i. We propose explicitly modeling them using neural networks, thus resulting in the estimated entropy of Equation 15.

Implementation: We will use Gaussian distributions for y_i^* and y_i and a Bernoulli distribution for ℓ_i. We enforce the constraint that σ_i^* and σ_i should be positive by applying the softplus activation function on these parameters. To summarise,

$$\begin{aligned} p(y_i^*|\mathbf{x}_i, \theta) &\sim \underbrace{\mathcal{N}(\mu_i^*, \sigma_i^{2*}|\mathbf{x}_i, \theta)}_{\text{True distribution of } y_i^*} \\ p(y_i|\mathbf{x}_i, \theta) &\sim \underbrace{\mathcal{N}(\mu_i, \sigma_i^2|\mathbf{x}_i, \theta)}_{\text{Dist. of observed values } y_i}\quad , \\ p(\ell_i|\mathbf{x}_i, \theta) &\sim \underbrace{\mathrm{Ber}(\lambda_i|\mathbf{x}_i, \theta)}_{\text{Distribution of } \ell_i}\quad . \end{aligned} \tag{16}$$

We model all these distributions with a single Bayesian neural network with stochastic parameters. The outputs of the Bayesian neural network are the parameters of the distributions $p(y_i^*|\mathbf{x}_i, \theta)$, $p(y_i|\mathbf{x}_i, \theta)$ and $p(\ell_i|\mathbf{x}_i, \theta)$, i.e. five outputs neurons for the set $\{\mu_i^*, \sigma_i^*, \mu_i, \sigma_i, \lambda_i\}$. The latter can then be used to compute the conditional entropy in Equation 15 as highlighted by the respective

[1] Note that this distribution is different from Eq. 2, where we fit a distribution for y using a Tobit likelihood.

Name	Num. features	Censorship	n_0	Acquisition size	Steps	repetitions	$\mathcal{D}^{\text{POOL}}$	$\mathcal{D}^{\text{VAL}}$	$\mathcal{D}^{\text{TEST}}$
Synthetic	1	44%	10	3	150	50	9000	250	500
BreastMSK	5	77%	5	3	150	50	1285	183	366
Metabric	9	42%	5	3	150	50	1523	76	305
Whas	6	58%	5	3	150	50	1310	65	263
GBSG	7	37%	5	3	150	50	1546	137	549
Support	14	32%	5	3	150	50	7098	355	1420
Churn	26	53%	5	3	150	50	1276	136	546
Credit Risk	47	30%	5	3	150	50	650	70	280
SurvMNIST	28×28	53%	100	5	100	25	60000	5000	5000

Table 1. Overview of the various datasets used in this analysis, including the number of features and the percentage of censorship in $\mathcal{D}^{\text{pool}}$. We also include n_0 as the initial data points in $\mathcal{D}^{\text{train}}$

colors. The parameters of the neural network, θ, are inferred using the total loss from the maximum likelihood estimation of all these distributions,

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathcal{C}}(\theta) + \mathcal{L}_{\text{GAUSS}}(\theta) + \mathcal{L}_{\text{BCE}}(\theta)\,. \tag{17}$$

Figure 3 shows the fit of the proposed model for all the different distributions on a synthetic dataset. Using all the explicit models of y_i^*, y_i and ℓ_i, we can compute the $\mathcal{C}$-BALD objective (Equation 13) and use it as an acquisition function in active learning.

5 Experiments

In this section, we present the results of the proposed acquisition function with multiple experiments on synthetic and real-world datasets. Source code for reproducing the experiments is available at: `https://github.com/fbohu/Censored-Active-Learning`.

Models: We implement the Bayesian Neural Network with stochastic parameters using Monte Carlo Dropout [7]. We use three layers, 128 hidden units, a dropout probability of 0.25, and the ADAM optimizer with a learning rate of $0.3 \cdot 10^{-3}$ [23] and the ReLU activation function[2].

Baselines: We compare the proposed acquisition function with the following baselines: **Random** acquisitions, which randomly acquires data points in $\mathcal{D}^{\text{pool}}$, the **Entropy** (Entropy) of Bayesian neural networks, which is proportional to variance between the individual's models in the sampled ensemble, $\text{Var}_{\theta \sim p(\theta|\mathcal{D})}[p(y_i|\mathbf{x}_i)]$, and the **BALD** objective from Equation 7.

Evaluation: To quantify the performance of the acquisition function, we evaluate the relative decrease in the area under the curve (RD-AUC) across the entire active learning experiment [41]. We compare the relative decrease to a baseline acquisition function (Random) and evaluate the models' right censored

[2] In Appendix C, we experiment with different model architectures.

negative log-likelihood (NLL) on a test set ($\mathcal{D}^{\text{test}}$). Since the NLL is not bounded by 0, we use the lowest NLL obtained across all the acquisition functions as a lower bound for the metric. We compute the average across all the number of acquisitions, N_{Acq}. The RD-AUC is defined as follows:

$$\text{RD-AUC} = \frac{1}{N_{\text{Acq}}} \sum_{i=0}^{N_{\text{Acq}}} \left(\frac{NLL_{\text{Random}} - NLL_s}{NLL_{\text{Random}}} \right) , \tag{18}$$

where NLL_s is the negative log-likelihood of the model with the acquisition function s and NLL_{Random} is the negative log-likelihood of from Random acquisition.

Synthetic Data: We begin our empirical evaluation of the proposed acquisition by considering the following 1D synthetic dataset, with $x_i = \mathcal{N}(5, 1)$, and,

$$y_i^* = \frac{1}{2}\sin(2x_i) + 2 + \varepsilon_i \qquad z_i = \frac{1}{2}\cos(2x_i) + 2 + \varepsilon_i \, , \tag{19}$$

$y_i = \min(y_i^*, z_i)$, and $\ell_i = \mathbb{1}\{y_i^* \leq z_i\}$ and $\varepsilon_i \sim \mathcal{N}(0, 0.01|x_i|)$. The dataset can be seen in Figure 2 and our proposed modeling fit in Figure 3. We generate a small pool of labelled data points ($n_0 = 10$), a larger set of unlabelled data points $|\mathcal{D}^{\text{pool}}| = 9000$, and a $|\mathcal{D}^{\text{test}}| = 500$. We train a model of the small pool of labeled data and acquire three new data points with labels every iteration. During each training step, we use a small validation set $\mathcal{D}^{\text{val}}$ with 250 observations to evaluate the models and apply early stopping on the right censored maximum likelihood.

Figure 3 shows the $\mathcal{C}$-BALD scores across the entire range of x. $\mathcal{C}$-BALD assigns a high mutual information value in regions where the censoring status changes, i.e. when the model is uncertain about the information that new samples will provide. In the right of Figure 3, we show the right censored negative log-likelihood for the different acquisition functions. We find that $\mathcal{C}$-BALD achieves the best overall fit of the data with the lowest NLL, which shows that it identifies which data point provides the most information to the model.

Real Datasets: We test the proposed functions on seven real-world datasets: five from a biomedical context [22] and two from a predictive analytics context [6]. Three datasets focus on estimating the survival time for various types of cancer patients (**BreastMSK**, **METABRIC**, and **GBSG**), one dataset for modeling the survival time of myocardial infarction (**WHAS**), and the last dataset estimates the survival time for critically-ill hospital patients (**SUPPORT**). For the predictive analytics datasets, we focus on predicting the time customers remain subscribed to a service (**Churn**) and the other on estimating the time for borrowers to repay their credit (**Credit Risk**)[3].

Table 1 summarises the datasets used in the experiments, including the number of features, the percentage of censored observations, and the total number of observations. Additionally, it includes a summary of the parameters used for the active learning experiments for each dataset. The results reported are averages

[3] A more extensive summary of these datasets can be found in Appendix B.3.

Dataset	Entropy	BALD	$\mathcal{C}$-BALD
Synthetic	8.65 ± 0.42	-0.12 ± 0.14	$\mathbf{33.49 \pm 1.11}$
BreastMSK	8.21 ± 1.43	-1.89 ± 0.66	$\mathbf{8.75 \pm 1.42}$
Metabric	-0.67 ± 0.39	2.25 ± 0.34	$\mathbf{18.26 \pm 0.94}$
whas	0.42 ± 0.27	$\mathbf{1.68 \pm 0.17}$	0.26 ± 0.32
GBSG	-0.81 ± 0.05	-0.04 ± 0.05	$\mathbf{5.58 \pm 0.05}$
support	0.70 ± 0.02	-0.53 ± 0.01	$\mathbf{4.55 \pm 0.02}$
churn	5.14 ± 0.31	0.17 ± 0.21	$\mathbf{32.75 \pm 0.87}$
credit risk	-0.72 ± 0.36	-0.17 ± 0.33	$\mathbf{22.11 \pm 0.64}$
Survmnist	-0.05 ± 0.28	1.06 ± 0.30	$\mathbf{13.47 \pm 0.66}$

Table 2. Relative decrease in the area under the curve (RD-AUC) compared to the Random scoring function. A higher value in the table represents better performance, with the best performance highlighted in **bold**.

over the number of repetitions for each dataset and acquisition function (mean ± standard error). Table 2 reports the RD-AUC compared across the different scoring functions. Figure 4 shows the right-censored NLL across the different runs for two real-world datasets. We find that the proposed acquisition function leads to better acquisition of new data points by obtaining a superior fit on the test set compared to the baselines.

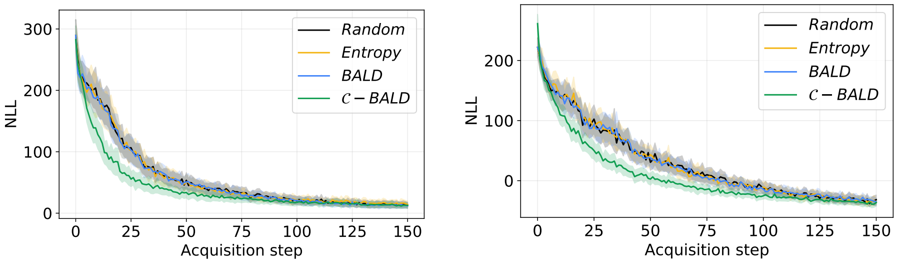

Fig. 4. Results of the real-world experiments on two of the seven datasets, namely the METABRIC and CREDIT RISK datasets, respectively. The figure shows the NLL (mean ± standard error) across the multiple repetitions of the experiment.

High-dimensional data: Lastly, we evaluate the performance of our proposed scoring functions with Bayesian convolutional neural networks on the **SurvMNIST** dataset [10]. In SurvMNIST, each label is replaced with a random draw from a Gamma distribution, with different distributional parameters across the labels [39]. The observations in the dataset are censored uniformly, between the minimum and the 90th percentile in the training set [10]. The initial training set contains ten samples from each class in the dataset[4]. The experiment on the SurvMNIST dataset shows that the proposed scoring functions outperformed the baseline functions, as shown in Table 2.

6 Related work

The study of the information that an experiment or observation provides was introduced by [31] and has often been the basis for new acquisition functions in active learning [34,33]. The study of information in censored experiments has traditionally focused on survival experiments, where observations are studied over time [14,16]. In survival experiments, an individual is observed for an amount of time and is considered censored if the person drops out of the experiment [15]. For the discrete and continuous case, the entropy calculations come down to the integral over the time an individual was observed [1], and entropy decreases after observations are censored [15]. In these and other settings with censored data, such as transportation systems [20,18], subscription-based businesses [5,3,36], and in health survival applications [37,30], data can be expensive to collect and label, necessitating the need for active learning in this context. Despite the challenges of censored data, there is limited research on active learning in this context. Two notable exceptions from the survival analysis literature include the work of [48], who proposed a query strategy based on discriminative gradients to identify the most informative points, and the work of [37], who suggested a query strategy for acquiring data points with the highest expected performance increase if their labels were known. A popular approach is Bayesian Active Learning with the BALD objective [17], specifically with its ability to work in conjunction with deep neural networks [8] and extensions to batch-acquisitions [25]. In the Deep Bayesian active learning, the BALD objective has primarily been used for classification tasks with MC Dropout models but has recently seen applications for deep regression tasks, such as estimating causal treatment effects [21] and for black-box models [24]. While plenty of research has focused on the BALD objective, to our knowledge, we are the first to explore the BALD objective in censored regression.

As previously mentioned, this work is motivated by predict-then-optimize scenarios, often involving dynamic feedback between supply and demand, which aligns with exploration/exploitation approaches as in Bayesian Optimization (BO) and Reinforcement Learning (RL), however in our work, we not focused on optimization, but increase the predictive quality of the censored regression models, used in these contexts. This is an important distiction, as supply is often limited and needs proper allocation, which is often done by optimizing a utility function, such as profit or cost [11]. While BO extensions for censored data (e.g. [19]) exist that handle uncertainty from censoring, focusing on active learning allows for the separation between learning from optimization, thus allowing for greater flexibility in applying and comparing different optimization techniques.

[4] The details of the gamma distributions and the model architecture can be found in Appendix B.4.

For instance, [11] shows that a Chance Constrained Mixed Integer Programming approach outperforms a BO baseline when censoring is not modeled. In contrast, while RL has been studied in the context of censored data (e.g., [47,9]), it is less suited to the type of problems targeted by our work due to the high cost of acquiring new data (e.g., deploying infrastructure). This restricts the extensive exploration typically needed in RL, making it impractical for real-world applications with tight exploration budgets.

7 Conclusion

This paper studies Bayesian active learning for censored regression problems. This problem is prevalent in many fields, such as engineering, marketing, finance, and medicine, where datasets often contain censored observations and obtaining new observations can be costly, thus constraining the learning process of the true underlying uncensored distribution and requiring careful strategic decision-making under uncertainty. Motivated by this challenge, we derive the entropy for censored distributions and propose the $\mathcal{C}$-BALD acquisition function, which accounts for censored observations. Empirically, across various synthetic and real datasets, we show that $\mathcal{C}$-BALD outperforms BALD on synthetic and real-world datasets.

References

1. Baxter, L.A.: A note on information and censored absolutely continuous random variables. Statistics & Risk Modeling **7**(1-2), 193–198 (1989)
2. Bickford Smith, F., Kirsch, A., Farquhar, S., Gal, Y., Foster, A., Rainforth, T.: Prediction-oriented bayesian active learning. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. pp. 7331–7348 (2023)
3. Chandar, P., Thomas, B., Maystre, L., Pappu, V., Sanchis-Ojeda, R., Wu, T., Carterette, B., Lalmas, M., Jebara, T.: Using survival models to estimate user engagement in online experiments. pp. 3186–3195 (2022)
4. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) **34**(2), 187–220 (1972)
5. Fader, P.S., Hardie, B.G.: How to project customer retention. Journal of Interactive Marketing **21**(1), 76–90 (2007)
6. Fotso, S., et al.: PySurvival: Open source package for survival analysis modeling (2019)
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of The 33rd International Conference on Machine Learning. pp. 1050–1059 (2016)
8. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1183–1192 (2017)
9. Goldberg, Y., Kosorok, M.R.: Q-learning with censored data. Annals of statistics **40**(1), 529 (2012)
10. Goldstein, M., Han, X., Puli, A., Perotte, A., Ranganath, R.: X-cal: Explicit calibration for survival analysis. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18296–18307 (2020)

11. Golsefidi, A.H., Hüttel, F.B., Peled, I., Samaranayake, S., Pereira, F.C.: A joint machine learning and optimization approach for incremental expansion of electric vehicle charging infrastructure. Transportation Research Part A: Policy and Practice **178**, 103863 (2023)
12. Haghighi, K., Ström, E.G., Agrell, E.: Sensing or transmission: Causal cognitive radio strategies with censorship. IEEE Transactions on Wireless Communications **13**(6), 3031–3041 (2014)
13. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
14. Hollander, M., Proschan, F., Sconing, J., STATISTICS, F.S.U.T.D.O.: Information in censored models. FSU Statistics Report M **701** (1985)
15. Hollander, M., Proschan, F., Sconing, J.: Measuring information in right-censored models. Naval Research Logistics (NRL) **34**(5), 669–681 (1987)
16. Hollander, M., Proschan, F., Sconing, J.: Information, censoring, and dependence. Lecture Notes-Monograph Series **16**, 257–268 (1990)
17. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning (2011)
18. Hüttel, F.B., Rodrigues, F., Pereira, F.C.: Mind the gap: Modelling difference between censored and uncensored electric vehicle charging demand. Transportation Research Part C: Emerging Technologies **153**, 104189 (2023)
19. Hutter, F., Hoos, H., Leyton-Brown, K.: Bayesian optimization with censored response data. arXiv preprint arXiv:1310.1947 (2013)
20. Hüttel, F.B., Peled, I., Rodrigues, F., Pereira, F.C.: Modeling censored mobility demand through censored quantile regression neural networks. IEEE Transactions on Intelligent Transportation Systems **23**(11), 21753–21765 (2022)
21. Jesson, A., Tigas, P., van Amersfoort, J., Kirsch, A., Shalit, U., Gal, Y.: Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2021)
22. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deep-surv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Research Methodology **18**(1), 24 (2018)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Kirsch, A.: Black-box batch active learning for regression. Transactions on Machine Learning Research (2023), expert Certification
25. Kirsch, A., van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
26. Kirsch, A., Gal, Y.: Unifying approaches in active learning and active sampling via fisher information and information-theoretic quantities. Transactions on Machine Learning Research (2022), expert Certification
27. Knaus, W., Harrell, F., Lynn, J., Goldman, L., Phillips, R., Connors, Jr, A., Dawson, N., Fulkerson, W., Califf, R., Desbiens, N., Layde, P., Oye, R., Bellamy, P., Hakim, R., Wagner, D.: The support prognostic model. objective estimates of survival for seriously ill hospitalized adults. study to understand prognoses and preferences for outcomes and risks of treatments. Annals of internal medicine **122**, 191–203 (1995)
28. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. vol. 30 (2017)

29. Lemeshow, S., May, S., Hosmer Jr, D.W.: Applied survival analysis: regression modeling of time-to-event data. John Wiley & Sons (2011)
30. Lian, J., Long, Y., Huang, F., Ng, K., Lee, F.M.Y., Lam, D.L., Fang, B.L., Dou, Q., Vardhanabhuti, V.: Imaging-based deep graph neural networks for survival analysis in early stage lung cancer using ct: A multicenter study. Frontiers in Oncology **12** (2022)
31. Lindley, D.V.: On a Measure of the Information Provided by an Experiment. The Annals of Mathematical Statistics **27**(4), 986 – 1005 (1956)
32. Liu, N., Ye, Y.: Humanitarian logistics planning for natural disaster response with bayesian information updates. Journal of Industrial & Management Optimization **10**(3), 901–919 (2014)
33. MacKay, D.J.C.: The evidence framework applied to classification networks. Neural Computation **4**(5), 720–736 (1992)
34. MacKay, D.J.C.: Information-Based Objective Functions for Active Data Selection. Neural Computation **4**(4), 590–604 (1992)
35. Mahajan, A.: Structure of optimal policies in active sensing. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5265–5268. IEEE (2012)
36. Maystre, L., Russo, D.: Temporally-consistent survival analysis. In: Advances in Neural Information Processing Systems (2022)
37. Nezhad, M.Z., Sadati, N., Yang, K., Zhu, D.: A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer. Expert Systems with Applications **115**, 16–26 (2019)
38. Osborne, M.A., Roberts, S.J., Rogers, A., Jennings, N.R.: Real-time information processing of environmental sensor network data using bayesian gaussian processes. ACM Transactions on Sensor Networks (TOSN) **9**(1), 1–32 (2012)
39. Pearce, T., Jeong, J.H., Jia, Y., Zhu, J.: Censored quantile regression neural networks for distribution-free survival analysis. In: Advances in Neural Information Processing Systems (2022)
40. Powell, J.L.: Censored regression quantiles. Journal of Econometrics **32**(1), 143–155 (1986)
41. Riis, C., Antunes, F., Hüttel, F.B., Azevedo, C.L., Pereira, F.C.: Bayesian active learning with fully bayesian gaussian processes. In: Advances in Neural Information Processing Systems (2022)
42. Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R.L., Rauschecker, H.F.: Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. J Clin Oncol **12**(10), 2086–2093 (Oct 1994)
43. Settles, B.: Active learning literature survey (2009)
44. Sharma, M., Farquhar, S., Nalisnick, E., Rainforth, T.: Do bayesian neural networks need to be fully stochastic? In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. vol. 206, pp. 7694–7722 (2023)
45. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. In: International Conference on Learning Representations (2018)
46. Tobin, J.: Estimation of Relationships for Limited Dependent Variables. Econometrica **26**(1), 24–36 (1958)
47. Tornede, A., Bengs, V., Hüllermeier, E.: Machine learning for online algorithm selection under censored feedback. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10370–10380 (2022)

48. Vinzamuri, B., li, Y., Reddy, C.: Active learning based survival regression for censored data. Proc. of the 2014 ACM International Conference on Information and Knowledge Management pp. 241–250 (2014)
49. Zhang, J., Huang, J., Wang, T., Zhao, J.: Dynamic optimization of emergency logistics for major epidemic considering demand urgency. Systems **11**(6), 303 (2023)