Variational Inference with Censored Gaussian Process Regressors

Andrea Karlova [1] **Rishabh Kabra** [2 1] **Daniel Augusto de Souza** [1] **Brooks Paige** [1]

Abstract

We consider the problem of Bayesian inference when some observations have been censored. In censored data, the dependent variable has been clipped, so we only know that the true value is at least as large (or as small) as the observation. Such data can be modeled using a Tobit likelihood, which can be viewed as a mixture between a normal distribution restricted on the domain without censoring treatment and a point mass at the boundary. This requires careful consideration when evaluating information-theoretic quantities, due to the mixed continuous and discrete probability measures. We introduce a novel approximate inference scheme for Gaussian process models with a Tobit likelihood, derive interpretable analytic expression for the Gaussian process evidence lower bound (ELBO) and demonstrate the resulting model's efficiency in learning Gaussian process posteriors for censored data relative to uncensored case.

1. Introduction

Measurements in the real world come from bounded-scale instruments or finite-time events. Such intervals require careful consideration at the boundaries, where the variable is clipped and the density ceases to be continuous. The Censored Gaussian aka Tobit likelihood is instrumental in modeling such data (Fig 2). While there has been recent work (Gammelli et al., 2022; Basson et al., 2023) advancing the inference of Censored Gaussian Processes, so far, most statistical and information-theoretic quantities of the likelihood distribution were not provided in the closed analytical form.

An appropriate likelihood assumption is essential to any latent modeling task. Choosing the wrong likelihood, such as the standard Gaussian, when it's impossible to observe data outside a given range, introduces systematic bias into the model. A misalignment between modeling assumptions and the empirical data distribution not only affects the model's predictive ability, but also its uncertainty quantification. This prese nts a two-fold challenge in fields such as medical diagnostics (Rao et al., 2016) or environmental monitoring (Friederichs & Hense, 2007) which not only care more about model uncertainty, but are also likelier to encounter censored or missing data (Chen et al., 2013).

The Tobit likelihood, despite being an appropriate choice when a censoring interval is fixed and given, has found limited use in Bayesian inference (Basson et al., 2023). A key missing piece is an inference procedure that would work in conjunction with expressive models such as Gaussian processes (GPs), without relying on high-variance Montecarlo estimation. While the posterior is analytically intractable (Ertin, 2007; Groot & Lucas, 2012), we provide feasible analytical approximation to the evidence lower bound, (Jaakkola & Jordan, 2000), which allows for more controllable inference, thus laying a more rigorous foundation for Bayesian inference with censored data.This enables the use of Variational Inference and gradient-based learning for GPs modeling censored data.

The main contribution of our paper is the derivation of the closed form formula for cross entropy of normal and censored normal distribution which we apply for formulating the interpretable closed-form evidence lower bound (ELBO) for Variational Inference when approximating the posterior corresponding to a Tobit likelihood. We demonstrate through GP regression experiments how a Tobit likelihood can be used as a plug-in replacement for the usual uncensored Gaussian.

2. Background and Related Work

We set out to infer a (latent) data generating function for censored data. Following a Bayesian approach, there are typically four choices to make: (i) a prior distribution $p(f)$ over the latent function space, (ii) a likelihood distribution $p(y_i|f)$ which specifies how the data would be generated given a latent function, (iii) an inference method for the posterior distribution $p(f|y_1, \ldots, y_N) \propto \prod_i p(y_i|f)p(f)$ given observed data, and (iv) a strategy to acquire new data points to aid learning. We go over each of these choices in

[1]University College London [2]Google DeepMind. Correspondence to: Andrea Karlova <a.karlova@ucl.ac.uk>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

the context of our problem, and review the relevant literature below.

2.1. A Likelihood for Censored Data

In a traditional regression problem, we observe a set of noise-corrupted output values, $\boldsymbol{y} \in \mathbb{R}^N$, which are dependent variables of a set of input values $\boldsymbol{X} \in \mathbb{R}^{N \times D}$. The goal is to recover the noise-free function $f(\boldsymbol{x})$. A typical choice is to assume a Gaussian likelihood model. However, if we know that not only is the data collection noisy but also clamped to the interval $[l, u]$, this gives rises to a different corruption process known as *censoring*. Mathematically, this process is:

$$y(\boldsymbol{x}) = \begin{cases} l, & \text{if } f(\boldsymbol{x}) + \varepsilon \leq l \\ f(\boldsymbol{x}) + \varepsilon, & \text{if } l < f(\boldsymbol{x}) + \varepsilon < u \\ u, & \text{if } f(\boldsymbol{x}) + \varepsilon \geq u \end{cases}, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma_y^2)$. Since this corruption corresponds to a measurable function, it also defines a distribution, known as the Tobit likelihood in reference to Tobin (Tobin, 1958). This approach has been extensively studied in econometrics (Amemiya, 1984; Robin, 2010) and survival analysis (Klein et al., 2003), with extensions to handle various censoring mechanisms and distributional assumptions, including machine learning (Pearce et al., 2022; Friederichs & Hense, 2007; Moradian et al., 2017).

Unfortunately, the Tobit likelihood is not absolutely continuous with respect to the Lebesgue measure; it lacks a defined probability density function (PDF), in contrast with the Gaussian likelihood. Nonetheless, the cumulative distribution function (CDF) exists as is defined as:

$$p(y \leq \xi \mid f(\boldsymbol{x}), l, u) = \begin{cases} 0, & \text{if } \xi < l \\ \Phi(\xi \mid f(\boldsymbol{x}), \sigma_y^2), & \text{if } l \leq \xi \leq u \\ 1, & \text{if } \xi > u \end{cases}, \tag{2}$$

where $\Phi(x \mid \mu, \sigma^2)$ is the CDF of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

2.2. Gaussian Processes (GPs)

After choosing a likelihood, another decision is which prior distribution over functions to use. The most common choice of prior in Bayesian active learning is the Gaussian process (GP) (Li et al., 2024). A GP is a simple, but flexible, distribution over functions built on the assumption that any two values $f(x)$ and $f(x')$ are correlated Gaussian random variables (Williams & Rasmussen, 2006). Given training and evaluation sets, $\boldsymbol{X}$ and $\boldsymbol{X}^*$, we can represent the output of the latent functions as $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)]$ and $\boldsymbol{f}^* = [f(\boldsymbol{x}_1^*), \ldots, f(\boldsymbol{x}_N^*)]$ with joint distribution:

$$p(\boldsymbol{f}, \boldsymbol{f}^*) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}^* \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_f & \boldsymbol{K}_{f*} \\ \boldsymbol{K}_{*f} & \boldsymbol{K}_* \end{bmatrix} \right), \tag{3}$$

where $[\boldsymbol{\mu}_f]_i = \mu(x_i)$ and $[\boldsymbol{K}_f]_{i,j} = k(x_i, x_j)$. The mean function $\mu(x)$ and kernel function $k(x, x')$ are the parameters of the Gaussian process distribution and specify the mean and covariance of the joint distributions of observations.

The reason for the popularity of GPs is due to their conjugacy with the Gaussian likelihood which allows not only the posterior distribution $p(\boldsymbol{f} \mid \boldsymbol{y})$ and predictive posterior distribution $p(\boldsymbol{f}^* \mid \boldsymbol{y})$ to be calculated exactly in closed form, but also the model evidence $p(\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{f})}[p(\boldsymbol{y} \mid \boldsymbol{f})]$ enabling gradient-based selection of model hyperparameters by maximizing the model evidence.

2.3. Variational Inference for GPs

For likelihoods other than Gaussian, the posterior distribution usually cannot be computed and requires the use of approximations. Possible strategies to deal with this include Markov Chain Monte Carlo (MCMC) methods (Neal, 1997), Laplace approximations (Williams & Barber, 1998; Barrett & Coolen, 2012), and variational inference techniques (Hensman et al., 2015). Due to its deterministic nature compared to MCMC, higher expressivity compared to Laplace approximation, and connections to scalable GPs for big data, variational inference (VI) became the default choice for GP approximations.

Variational inference works by defining an approximate posterior $p(\boldsymbol{f}, \boldsymbol{f}^*) \approx q(\boldsymbol{f}, \boldsymbol{f}^*)$ where the posterior of the training data is a Gaussian distribution with learnable parameters $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f} \mid \hat{\boldsymbol{f}}, \boldsymbol{A})$ and the conditional predictive posterior is simply $q(\boldsymbol{f} \mid \boldsymbol{f}^*) = p(\boldsymbol{f}^* \mid \boldsymbol{f})$. The parameters of the approximation are trained by maximizing the evidence lower bound (ELBO):

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}[\log p(\boldsymbol{y} \mid \boldsymbol{f})] - KL[q(\boldsymbol{f}) \,||\, p(\boldsymbol{f})], \tag{4}$$

with the property that maximizing the ELBO jointly maximizes the evidence, allowing for hyperparameter optimization, and minimizes the divergence between the approximated posterior and the true posterior, reducing the error of the approximation. This property is one of the appealing factors of variational inference.

3. Our Method

3.1. Mixed Continuous and Discrete Measures

Censoring is a data treatment with the following mechanism: events which happen outside our observation window are

assigned the boundary value of this window. This does not change our perspective on how often the events occur (unlike truncation). But it changes the odds of how often events occur at the boundaries. Say we observe data from a normal distribution with values censored to lie in $[l, u]$. For an underlying distribution with density $\mathcal{N}(x; \mu, \sigma)$, we write the mixed density:

$$p(x;\mu,\sigma,l,u) = \Phi(\tfrac{l-\mu}{\sigma})\delta_{\{l\}}(x) + \mathcal{N}(x;\mu,\sigma)\mathbb{I}_{(l,u)} + [1-\Phi(\tfrac{u-\mu}{\sigma})]\delta_{\{u\}}(x), \tag{5}$$

where $\mathbb{I}_{(l,u)}$ is the identificator of belonging into the observation window (l, u) and $\delta_{\{z\}}(x)$ is the Dirac delta function, having mass at $x = z$ and null otherwise. The Dirac delta terms capture the size of the censored tails of the underlying density (see Fig 2).

When working with the censored probability measure, the usual framework which holds either for purely discrete or purely continuous probability distributions cannot be applied directly. We provide an exhaustive introduction to the problem in Appendix B. Essentially, we need to separately treat the areas where the probability density is defined to those with the discontinuities caused by mixing in the point masses of discrete distribution. While (Nair et al., 2006) denotes the mixed distribution by a mixed pair and considers the entropy of a mixed pair of discrete and continuous distribution in connection with entropy rate of MCMC, we carefully treat the problem using basic measure theory (Kallenberg, 2021).

3.2. Censored Regression

With the censored targets, the regression problem can be reformulated as: $y_i = f_i + \epsilon_i$, where $i = 1, \ldots, n$, and $\epsilon_i \sim \mathcal{N}_c(\varepsilon|0, \sigma_y^2, l, u)$ comes from a censored normal distribution with the censoring thresholds $l < u$. We can see this problem also as a combination of two probit models placed on the lower threshold $p(y_i^l = 1|\boldsymbol{x}) = \Phi(f(\boldsymbol{x})|l, \sigma_y^2)$, and upper threshold $p(y_i^u = 1|\boldsymbol{x}) = 1 - \Phi(f(\boldsymbol{x})|u, \sigma_y^2)$, while having a normal regression model $p(\boldsymbol{y}|\boldsymbol{f}, \sigma_y) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2\boldsymbol{I}_{nn})$ otherwise.

From this initial motivation, we can write the likelihood $p(\boldsymbol{y}|\boldsymbol{f})$ of the censored model as:

$$\prod_{y_i \le l} \Phi(l|f_i, \sigma_y^2) \prod_{l<y_i<u} \mathcal{N}(y_i|f_i, \sigma_y^2) \prod_{y_i \ge u} [1 - \Phi(u|f_i, \sigma_y^2)]. \tag{6}$$

We place a GP prior over the latent function $\boldsymbol{f}$ and obtain the following posterior distribution over the latent parameters $\boldsymbol{f}$:

$$p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\theta}) = \frac{\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{nn})}{p(\mathcal{D}|\boldsymbol{\theta})} p(\boldsymbol{y}|\boldsymbol{f}), \tag{7}$$

where we denote the observed data as $\mathcal{D} = \{(\boldsymbol{x_i}, y_i), i = 1 \ldots n\}$ and $\boldsymbol{y} = (y_1, \ldots y_n)$ is the vector of labels.

3.3. ELBO for Censored Regressor

As the marginal likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ is not analytically tractable, we assume a Gaussian approximation to the posterior distribution, parametrised by $\hat{\boldsymbol{f}}$ and $\boldsymbol{A}$:

$$p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\theta}) \approx q(\boldsymbol{f}|\mathcal{D}, \hat{\boldsymbol{\theta}}) = \mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}}, \boldsymbol{A}), \tag{8}$$

and minimize the KL divergence between the variational approximation of the posterior and the posterior itself. We write the KL divergence in terms of the censored normal likelihood $p(\boldsymbol{y}|\boldsymbol{f})$ and Gaussian prior $p(\boldsymbol{f})$ as:

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) \ge \sum_{i=1}^{n} \mathbb{E}_{q(f_i|\mathcal{D},\hat{\boldsymbol{\theta}})} \ln p(y_i|f_i) - - KL\left[q(\boldsymbol{f}|\mathcal{D}, \hat{\boldsymbol{\theta}})||p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})\right] \equiv ELBO \tag{9}$$

The derivation of the cross-entropy is presented in Appendix C.1. The mean and scale parameters of the variational posterior q are denoted $\hat{f}$ and a_{ii}. Taking the final form of the cross-entropy from Eq (22), we arrive at the closed form version of the ELBO:

$$\begin{aligned}
\ln & p(\mathcal{D}|\boldsymbol{\theta}) \ge \\
& - \sum_{i=1}^{n} \log[\sqrt{2\pi}\sigma_y] \left[\Phi\left(\tfrac{u-\hat{f}_i}{a_{ii}}\right) - \Phi\left(\tfrac{l-\hat{f}_i}{a_{ii}}\right)\right] \\
& + \sum_{i=1}^{n} \tfrac{(y_i-\hat{f}_i)^2+a_{ii}^2}{2\sigma_y^2} \left[\Phi\left(\tfrac{u-\hat{f}_i}{a_{ii}}\right) - \Phi\left(\tfrac{l-\hat{f}_i}{a_{ii}}\right)\right] \\
& + \sum_{i=1}^{n} \left[\log \Phi\left(\tfrac{y_i-u}{\sigma_y}\right) + \tfrac{(u+\hat{f}_i-2y_i)a_{ii}}{2\sigma_y^2}\right] \mathcal{N}\left(\tfrac{u-\hat{f}_i}{a_{ii}}\right) \\
& + \sum_{i=1}^{n} \left[\log \Phi\left(\tfrac{l-y_i}{\sigma_y}\right) - \tfrac{(l+\hat{f}_i-2y_i)a_{ii}}{2\sigma_y^2}\right] \mathcal{N}\left(\tfrac{l-\hat{f}_i}{a_{ii}}\right) \\
& - \frac{1}{2}\left[\ln|\boldsymbol{A}| - \ln|\boldsymbol{K_{nn}}| - n\right] \\
& - \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{K_{nn}^{-1}A}) + \hat{\boldsymbol{f}}^T \boldsymbol{K_{nn}^{-1}} \hat{\boldsymbol{f}}\right].
\end{aligned} \tag{10}$$

The first two terms of the ELBO corresponds to the cross-entropy of the two uncensored Gaussian distributions less one half and scaled by variational approximation of the probability mass of the observation window between l and u; the third term captures the contribution of the upper censoring limit u while the fourth term corresponds to the contribution at l. The logarithm of normal CDF times the normal density is the contribution of the atomic mass at the boundary while the second linear term in the equation describes how far we are from the exact fit. It is easy to see that if l and u are set to $-\inf$ and $\inf$, then the second and third terms vanish (because $\mathcal{N}(x)$ becomes 0). The CDF difference in the first line reduces to 1.0, and the whole

Table 1: **Metrics evaluating 1D regression.** The subscript c denotes a metric computed on censored datapoints only.

	Censored Gaussian				Uncensored Gaussian			
	MAE	MAE_c	NLPD	$NLPD_c$	MAE	MAE_c	NLPD	$NLPD_c$
left-censored	0.40	0.14	7.87	15.74	0.67	0.59	2.03	1.93
right-censored	0.61	0.00	549.95	1,075.87	0.62	0.09	56.41	2.77
symmetric-censored	0.40	0.34	37.44	41.56	0.59	0.55	122.10	96.69

Table 2: **Metrics on real-world regression tasks**.

	Censored Gaussian		Uncensored Gaussian	
	MAE	NLPD	MAE	NLPD
Credit Risk: (40% test data)	9.5663	25.3787	9.5690	11.1385
gbsg cancer: (given test set)	21.1881	17.8265	19.8403	42.9532

ELBO reduces to the ELBO of an uncensored Gaussian. The last two terms corresponds to KL divergence between Gaussian prior and Gaussian variational approximation.

4. Experiments

4.1. GP Regression using Variational Inference

In Fig 1, we compare three ways of learning a synthetic function underlying censored observations. We start with uncensored Gaussian likelihood, a default likelihood for GP regression, and assumes no knowledge of the censoring. We use Variational Inference to fit the model. The second model is Censored Gaussian fit via Montecarlo simulation: this method samples from the (Multivariate Normal) variational posterior to parameterize the Censored Gaussian likelihood. We take 10,000 samples to facilitate convergence. The last model we consider is Censored Gaussian fit via Variational Inference: we use the posterior parameters to compute a closed-form expected log-likelihood as part of the ELBO, as derived in (10).

All methods use an RBF kernel with the same initialization, with a full matrix (Cholesky parameterized) covariance. We use L-BFGS (Liu & Nocedal, 1989) for optimization.

Fig 1 shows the benefit of taking censoring into account. The Uncensored Gaussian likelihood matches the empirical data density but completely fails to recover the true structure beyond the censoring bounds. It exhibits very low uncertainty around the censored regions due to the accumulated data density. The Censored Gaussian likelihood shows a significant improvement in capturing the structure of the generative function. Even with MC simulation from the approximate posterior, we achieve an almost unbiased fit. The caveat is the method is high-variance and very inefficient. We achieve the best fits using Variational Inference at significantly lower sample-complexity.

Metrics. To compare the uncensored and censored fits quantitatively, we compute the following 2 metrics on test-case data: Mean Absolute Error (MAE) and Negative Log Predictive Density (NLPD). MAE corresponds to assuming a common Laplace predictive likelihood (with scale 1). It is unambiguous and fair to both the uncensored and censored Gaussian likelihood models. For NLPD we evaluate both the uncensored and censored posteriors under the *censored* predictive likelihood—albeit with the respective noise parameters. In simple terms, this corresponds to passing the uncensored likelihood model through a clamping function.

Real-World Regression Tasks. We tested our model (10) on two real-life dataset with naturally censored dependent variables: the GBSG Cancer dataset (Katzman et al., 2018) and a Credit Risk dataset[1]. The results in Table 2 show that the Censored Gaussian is a drop-in replacement for the Uncensored likelihood when trained with our ELBO.

5. Discussion

We've shown that for GP Censored Regression it is in fact possible to derive tractable objective functions and validated our findings empirically by comparing a Censored Gaussian Regressor with the default Uncensored likelihood and comparing Gaussian processes fit using our derived ELBO versus a MC fit. Our derivations provide a general template to handle mixtures of discrete and continuous measures, or measures restricted to specific domains. Our work enables the adoption of more realistic modeling assumptions, due to its interpretability, which reflect how we observe the real world in practice.

[1] https://github.com/square/pysurvival/raw/master/pysurvival/datasets/credit_risk.csv

References

Amemiya, T. Tobit models: A survey. *Journal of econometrics*, 24(1-2):3–61, 1984.

Barrett, J. and Coolen, A. Gaussian process regression for survival analysis with interval censored data. *Biometrika*, 99(1):1–11, 2012.

Basson, M., Louw, T., and Smith, T. Variational tobit gaussian process regression. *Statistics and Computing*, 33(3), March 2023. ISSN 0960-3174. doi: 10.1007/s11222-023-10225-3.

Chen, N., Qian, Z., Nabney, I. T., and Meng, X. Wind power forecasts using gaussian processes and numerical weather prediction. *IEEE Transactions on Power Systems*, 29(2): 656–665, 2013.

Ertin, E. Gaussian process models for censored sensor readings. pp. 665–669, Los Alamitos, CA, USA, aug 2007. IEEE Computer Society. doi: 10.1109/SSP.2007.4301342.

Friederichs, P. and Hense, A. Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135(6):2365–2378, June 2007. ISSN 0027-0644. doi: 10.1175/mwr3403.1. URL http://dx.doi.org/10.1175/MWR3403.1.

Gammelli, D., Rolsted, K. P., Pacino, D., and Rodrigues, F. Generalized multi-output gaussian process censored regression. *Pattern Recognition*, 129:108751, 2022.

Groot, P. and Lucas, P. J. Gaussian process regression with censored data using expectation propagation. 2012.

Hensman, J., Matthews, A., and Ghahramani, Z. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pp. 351–360. PMLR, 2015.

Jaakkola, T. S. and Jordan, M. I. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10, 2000. doi: 10.1023/A:1008932416310.

Kallenberg, O. *Foundations of Modern Probability*, volume 1. Springer, 3 edition, 2021.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), February 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1. URL http://dx.doi.org/10.1186/s12874-018-0482-1.

Klein, J. P., Moeschberger, M. L., et al. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.

Li, Y. L., Rudner, T. G. J., and Wilson, A. G. A study of bayesian neural network surrogates for bayesian optimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Moradian, H., Larocque, D., and Bellavance, F. Survival forests for data with dependent censoring. *Statistical Methods in Medical Research*, 28(2):445–461, August 2017. ISSN 1477-0334. doi: 10.1177/0962280217727314. URL http://dx.doi.org/10.1177/0962280217727314.

Nair, C., Prabhakar, B., and Shah, D. On entropy for mixtures of discrete and continuous variables. *arXiv preprint cs/0607075*, 2006.

Neal, R. M. Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*, 1997.

Pearce, T., Jeong, J.-H., jia, y., and Zhu, J. Censored quantile regression neural networks for distribution-free survival analysis. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7450–7461. Curran Associates, Inc., 2022.

Rao, A., Monteiro, J., and Mourao-Miranda, J. Prediction of clinical scores from neuroimaging data with censored likelihood gaussian processes. In *2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pp. 1–4. IEEE, 2016.

Robin, J.-M. *Tobit model*, pp. 323–328. Palgrave Macmillan UK, London, 2010. ISBN 978-0-230-28081-6. doi: 10.1057/9780230280816_35. URL https://doi.org/10.1057/9780230280816_35.

Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pp. 24–36, 1958.

Williams, C. K. and Barber, D. Bayesian classification with gaussian processes. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1342–1351, 1998.

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

A. Experiments

We present figure with the results[2] from the investigation of the 1D regression. We fit the underlying generative function, $y = (5x - 10)\sin(10x - 20) + \epsilon$ using $n = 15$ data points. These are sampled and censored differently. We consider case of left censoring at bound -3, right censoring at bound 0 and symmetric censoring with bounds -2 and 2.

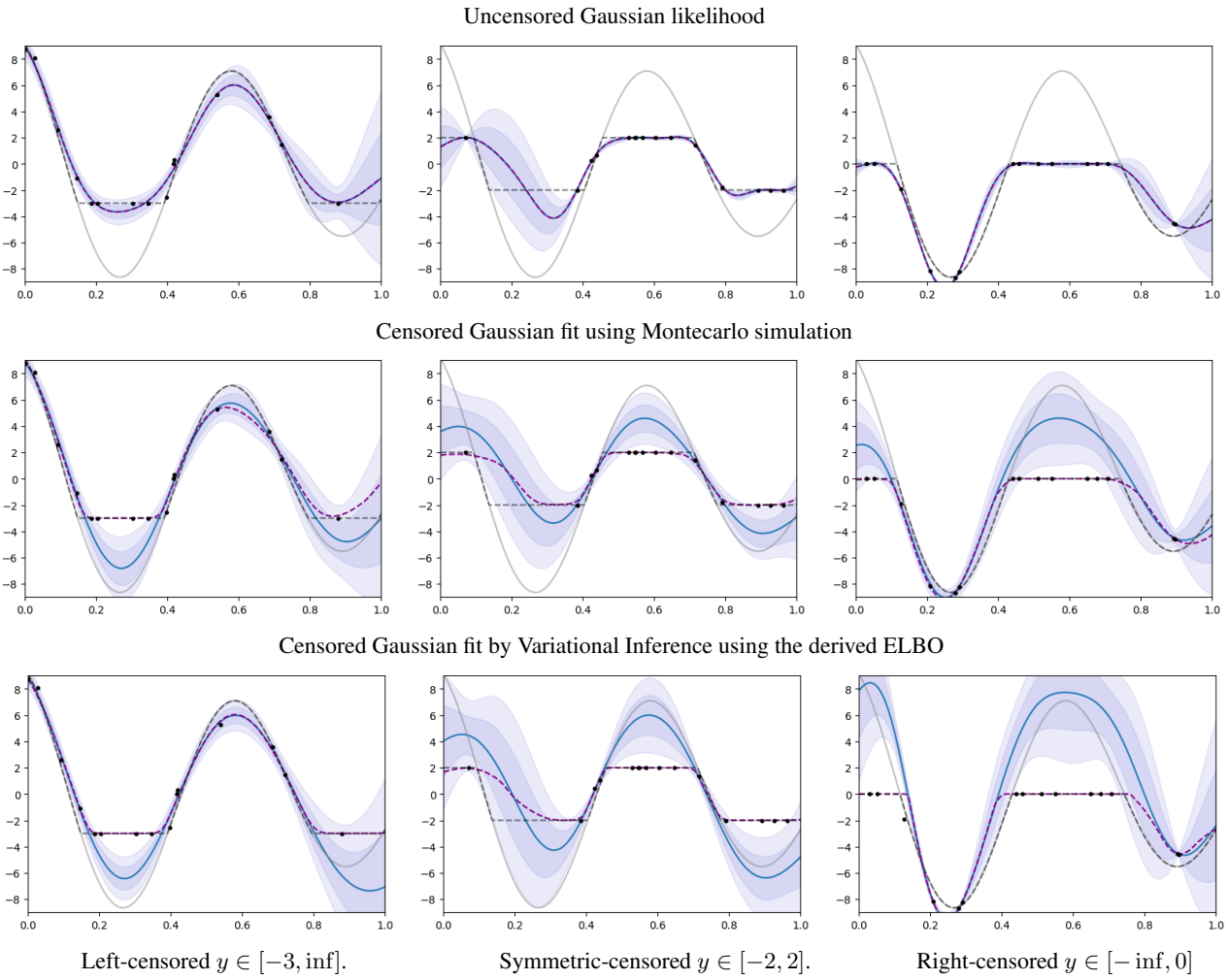

Figure 1: **One-dimensional GP regression.** We fit the underlying generative function, $y = (5x - 10)\sin(10x - 20) + \epsilon$ using n=15 data points. These are sampled and censored differently per column. The solid blue and gray lines show the uncensored prediction (latent posterior mean) and ground-truth function respectively. The purple and gray dashed lines show the censored predictive posterior mean and censored ground-truth respectively. The regions shaded in dark and light blue show the [15%, 85%] and [2.5%, 97.5%] percentiles of the latent posterior respectively.

B. Censoring and Truncating

Let us handle the missing data problem by either truncating or censoring the data. In both cases we can assume that the data comes from some latent generative process. The observations are revealed to us only when their values are within the observation window. When truncating the data, we discard the records when the outcome of the measurement falls outside the observation window. This makes us unaware of anything happening outside the observation boundaries including the quantity of the data we drop. On the other hand, during the censoring process, the values outside the observation window are assigned an artificial threshold value. So we keep the additional information about the tails by counting the occurrence

[2] https://sites.google.com/view/ml-with-censored-data

of possible values outside our observation boundaries. The censoring is often induced by the sensitivity of measuring instruments or the information availability when the value is within a certain bounds. Both data treatments inevitably impacts our modeling choices: the probability distribution of the data which we observe is different to the distribution of the underlying latent data which we could observe if we had a better measuring equipment.

In the following we demonstrate how the operation of data censoring and truncation impacts the underlying probability measures. Consider a diffusion measure ν_{df} that is supported on $\mathbb{R}$. Say the range of the values we have access to is restricted to some compact interval $[l, u]$. This naturally impacts our perception of the odds with which we measure particular events. So the first conclusion we make is that we observe the restriction of measure ν_{df} on the interval $[l, u]$. Denote the measure of the data distribution observed by us as ν_{emp}. From our view point, the values which we cannot observe due to the restriction are considered as impossible and these events have a zero probability. This implies that the supports of the measures ν_{emp} and ν_{df} differ: $supp_{\nu_{emp}} \subset supp_{\nu_{df}}$. We further conclude, that the measures are not equivalent as their domains of impossible events differ.

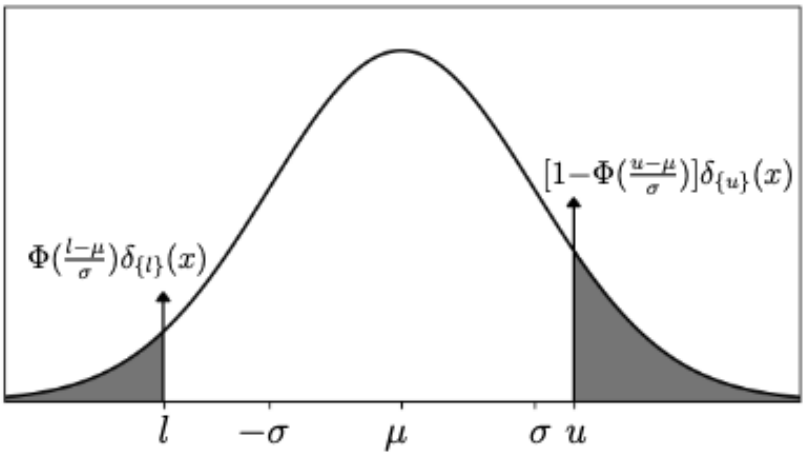

Figure 2: **A Tobit likelihood aka Censored Gaussian.** The support is restricted to the interval $x \in [l, u]$. Analytically this is handled by placing point masses at $x = l$ and $x = u$. Despite the non-existence of the pdf at these points, we show how to derive the distribution's entropy and cross-entropy.

Let us associate a normal distribution with density $\phi(x; \mu, \sigma)$ with our base distribution ν_{df}. When we choose to discard the records outside our observation range, the values appear to us to be more frequent. This impacts the probability measure by changing its scale comparing to the original distribution. For example, for the underlying normal distribution with density $\phi(x; \mu, \sigma)$ we have:

$$c \int_l^u \phi(x; \mu, \sigma) \mathrm{d}x = c \left[\Phi(\tfrac{u-\mu}{\sigma}) - \Phi(\tfrac{l-\mu}{\sigma}) \right] = 1, \tag{11}$$

and we simply infer the required scaling constant c. The truncated probability measure ν^t_{emp} is given by:

$$\nu^t_{emp} \equiv 0 \cdot \nu_{df}\delta_{(-\infty,l)} + \tfrac{1}{C}\nu_{df}\delta_{[l,u]} + 0 \cdot \nu_{df}\delta_{(u,\infty)}, \text{ with } C \equiv \Phi(\tfrac{u-\mu}{\sigma}) - \Phi(\tfrac{l-\mu}{\sigma}). \tag{12}$$

When censoring, the events which happens outside our observation domain are assigned the boundary value of the domain. This does not impact our perspective on how often the events occur. Nevertheless, it changes the odds of how often occur the events on the boundaries. We need to modify the latent measure ν_{df} restricted on the interval $[l, u]$ by placing the two atomic measures on its boundaries, where the mass of the atoms corresponds to the masses of the tails of the base latent measure:

$$\nu^c_{emp} \equiv 0 \cdot \nu_{df}\delta_{(-\infty,l)} + \Phi(\tfrac{l-\mu}{\sigma}) \cdot \nu_{df}\delta_l + \nu_{df}\delta_{(l,u)} + [1 - \Phi(\tfrac{u-\mu}{\sigma})] \cdot \nu_{df}\delta_u + 0 \cdot \nu_{df}\delta_{(u,\infty)}. \tag{13}$$

These changes to the probability distribution introduce a bias to the standard modeling approaches. For example, if we consider fitting linear regression to the censored or truncated data, the OLS estimator becomes biased and inconsistent. Figure 1 demonstrates the bias introduced to the OLS fit with the true values $\beta_0 = 1$, $\beta_1 = 1$ and $\sigma = 2$ of the parameters

of the underlying generative process $f(x) = \beta_0 + \beta_1 x + \sigma\varepsilon^2$ and $\varepsilon \sim N(0, 1)$. When we consider the Bayesian linear model, choosing the wrong likelihood distribution will result in introducing the bias too. Figure 2 demonstrates the impact of placing the wrong probability distribution over the likelihood of the modified data. When placing the censored normal distribution over the likelihood of the censored data, we recover the correct estimates of the parameters of the underlying generative process. The correction for the linear regression model, which removes the bias of OLS estimators are available in the literature. Correcting the biases for more complex high-dimensional models is yet to be systematically explored.

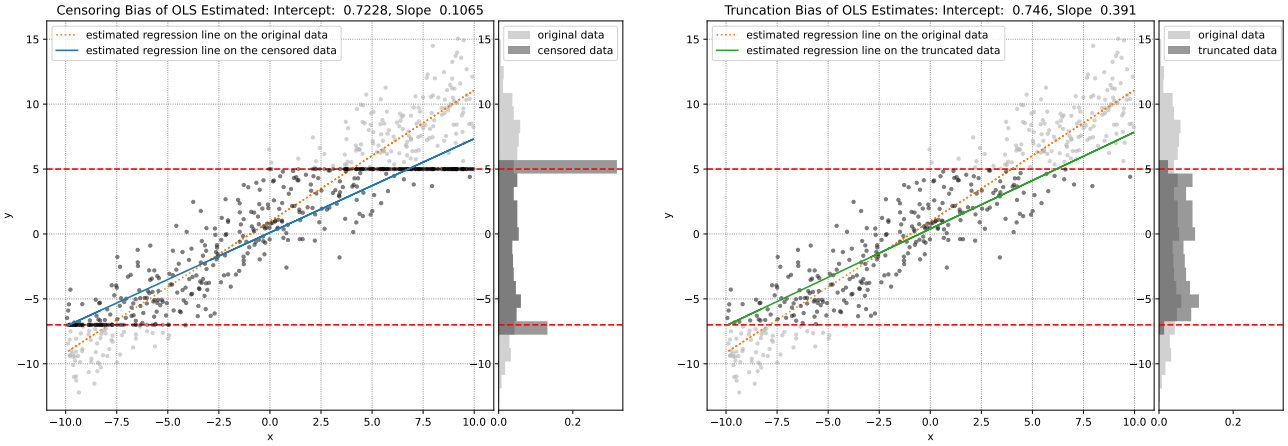

Figure 3: The bias of the OLS estimator fitted to the censored and truncated data. The original data are sampled from the regression line $f(x) = 1 + x + 2\varepsilon^2$ where $\varepsilon \sim N(0, 1)$.

C. Cross Entropy of Censored Normal

Using the measure defined in (13) we derive the cross entropy of the censored normal distribution. The cross entropy of ν wrt to reference measure ρ defined on some measurable space $(\Omega, \mathcal{A})$, with ν being absolutely continuous wrt to a reference measure ρ, is defined as:

$$H\left(\frac{\mathrm{d}\nu}{\mathrm{d}\rho}\right) = \int_{\Omega} \log \frac{\mathrm{d}\nu}{\mathrm{d}\rho} \rho(\omega), \tag{14}$$

where $\mathrm{d}\nu/\mathrm{d}\rho$ corresponds to the Radon-Nikodym derivative of the two considered measures.

Recall from section B that the latent diffusive measure ν_{df} is associated with the Lebesque measure with mixed-in Dirac measures which guarantees the existence of the density. The Radon-Nikodym decomposition of the measures characterises the decomposition of the measure into its diffusive part and the discrete part, i.e. allows to mix in the atoms which corresponds to the singular measures wrt Lebesque measure and atomic mesures without jeopardizing the existence of the integral (14). This does not pose any issue for the existence of the cross-entropy as long as the absolute continuity of measure ν wrt ρ is satisfied. We define the reference measure as $\rho = \lambda + \delta_l + \delta_u$, where λ denotes the Lebesque measure. The measure ν^c_{emp} is absolutely continuous wrt ρ as anytime $\rho(A) = 0$, then also $\nu^c_{emp}(A) = 0$ for any $A \in \mathcal{A}$. The opposite does not hold and so the measures ν^c_{emp} and ν are not equivalent. The Radon-Nykodym derivative of $\frac{\mathrm{d}\nu^c_{emp}}{\mathrm{d}\rho}$ defines the density $p(\mu, \sigma, l, u)$, where:

$$p(x; \mu, \sigma, l, u) = \Phi(\tfrac{l-\mu}{\sigma})\delta_l(x) + \phi(x; \mu, \sigma)\mathbb{I}_{(l,u)}(x) + [1 - \Phi(\tfrac{u-\mu}{\sigma})]\delta_u(x). \tag{15}$$

KL-divergence for the probability measures ν_1, ν_2, where ν_2 is absolutely continuous wrt to ν_1 and both measures are absolutely continuous wrt to ρ, is defined as follows:

$$D\left(\frac{\mathrm{d}\nu_1}{\mathrm{d}\rho} \middle\| \frac{\mathrm{d}\nu_2}{\mathrm{d}\rho}\right) = \int_{\Omega} \left[\log \frac{\mathrm{d}\nu_2}{\mathrm{d}\nu_1}\right] \frac{\mathrm{d}\nu_1}{\mathrm{d}\rho} \rho(\omega) \tag{16}$$

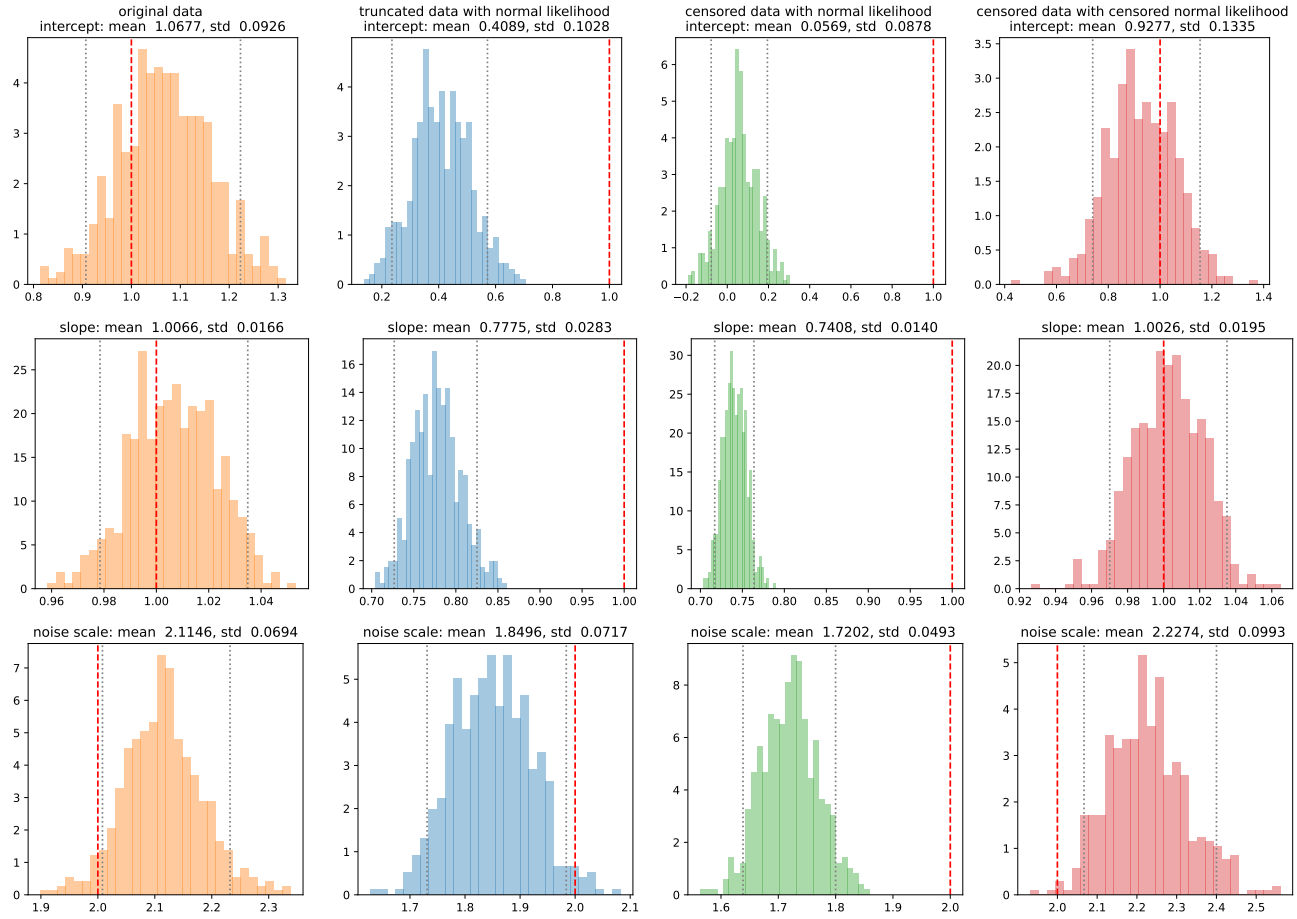

Figure 4: The posterior distribution of the parameters of the estimated linear model fitted to the dataset 1) without any data treatment, 2) with applying the truncation, 3) applying the the censoring. The red vertical line indicates the true value of the parameter used to generate the data. The dotted lines corresponds to 5% and 95% quantile.

This implies that we can only compute one-sided KL divergence between normal and censored normal probability distribution. We cannot swap the sides and the symmetrisation of the KL-divergence between normal and censored normal probability distribution is not possible. This is because the censored normal distribution is absolutely continuous wrt to the normal distribution. The normal distribution is not absolutely continuous wrt to the censored normal as it assigns non-zero mass to the tails, which the censored normal measures as a null set. This also provides guidance on how to compute KL-divergence between two censored normal distributions with different censoring bounds. The cross-entropy is the KL-divergence less the entropy of left-hand side probability measure, so the same rules applies to satisfy the existence of the integral.

C.1. Derivation of the Cross Entropy

Let us derive the (negative) cross entropy of the normal distribution with the density $\phi(m,s)$ and censored normal distribution $p(\mu,\sigma;l,u)$ with the underlying normal density $\phi(\mu,\sigma)$:

$$\begin{aligned}
&\mathbb{E}_{\phi(m,s)} \log p(l,u) \\
&\quad= \int_{\mathbb{R}} \phi(x;m,s) \log \left[\Phi(\tfrac{l-\mu}{\sigma})\delta_l(x) + \phi(x;\mu,\sigma)\delta_{(l,u)}(x) + [1-\Phi(\tfrac{u-\mu}{\sigma})]\delta_u(x)\right] \mathrm{d}x \\
&\quad= \left[\phi(x;m,s) \log \left[\Phi(\tfrac{l-\mu}{\sigma})\delta_l(x) + \phi(x;\mu,\sigma)\delta_{(l,u)}(x) + [1-\Phi(\tfrac{u-\mu}{\sigma})]\delta_u(x)\right]\right]_{x=l} \\
&\qquad+ \int_l^u \phi(x;m,s) \log \left[\Phi(\tfrac{l-\mu}{\sigma})\delta_l(x) + \phi(x;\mu,\sigma)\delta_{(l,u)}(x) + [1-\Phi(\tfrac{u-\mu}{\sigma})]\delta_u(x)\right] \mathrm{d}x \\
&\qquad+ \left[\phi(x;m,s) \log \left[\Phi(\tfrac{l-\mu}{\sigma})\delta_l(x) + \phi(x;\mu,\sigma)\delta_{(l,u)}(x) + [1-\Phi(\tfrac{u-\mu}{\sigma})]\delta_u(x)\right]\right]_{x=u} \\
&\quad= \phi(\tfrac{l-m}{s}) \log[\Phi(\tfrac{l-\mu}{\sigma})] + \int_l^u \phi(x;m,s) \log \phi(x;\mu,\sigma) \mathrm{d}x + \phi(\tfrac{u-m}{s}) \log[\Phi(\tfrac{\mu-u}{\sigma})].
\end{aligned} \tag{17}$$

Let us compute the middle term which corresponds to the cross-entropy of two unscaled normal distributions restricted to the compact interval $[l,u]$:

$$\begin{aligned}
I &\equiv \int_l^u \phi(x;m,s) \log \phi(x;\mu,\sigma) \mathrm{d}x \\
&= -\tfrac{1}{2}\log[2\pi\sigma^2] \int_l^u \phi(x;m,s) \mathrm{d}x \; - \tfrac{1}{\sqrt{2\pi s^2}} \int_l^u \tfrac{(x-\mu)^2}{2\sigma^2} e^{-\frac{(x-m)^2}{2s^2}} \mathrm{d}x.
\end{aligned} \tag{18}$$

The first term is the difference of the normal CDF at the transformed boundaries. For computing the second term we use the following trick: $[(x-m)+(m-\mu)]^2 = (x-m)^2 + 2x(m-\mu) + \mu^2 - m^2$:

$$\tfrac{1}{\sqrt{2\pi s^2}} \int_l^u \tfrac{(x-\mu)^2}{2\sigma^2} e^{-\frac{(x-m)^2}{2s^2}} \mathrm{d}x = \tfrac{s^2}{\sigma^2\sqrt{2\pi s^2}} \int_l^u \tfrac{(x-m)^2 + 2x(m-\mu) + \mu^2 - m^2}{2s^2} e^{-\frac{(x-m)^2}{2s^2}} \mathrm{d}x. \tag{19}$$

To evaluate the first term in (19) we use substitution $z = \frac{x-m}{s}$. Because $\left[-ze^{-\frac{z^2}{2}}\right] \mathrm{d}z = z^2 e^{-\frac{z^2}{2}} - e^{-\frac{z^2}{2}}$, we can compute the second term of the above integral in the closed form:

$$\begin{aligned}
\tfrac{1}{\sqrt{2\pi}} \int_{\frac{l-m}{s}}^{\frac{u-m}{s}} z^2 e^{-\frac{z^2}{2}} \mathrm{d}z &= \tfrac{1}{\sqrt{2\pi}} \int_{\frac{l-m}{s}}^{\frac{u-m}{s}} e^{-\frac{z^2}{2}} \mathrm{d}z - \tfrac{1}{\sqrt{2\pi}} \left[ze^{-\frac{z^2}{2}}\right]_{\frac{l-m}{s}}^{\frac{u-m}{s}} \\
&= \Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s}) - \tfrac{u-m}{s}\phi(\tfrac{u-m}{s}) + \tfrac{l-m}{s}\phi(\tfrac{l-m}{s}).
\end{aligned} \tag{20}$$

Let us evaluate the middle term in (19). Recall $\left[-e^{-\frac{z^2}{2}}\right] \mathrm{d}z = ze^{-\frac{z^2}{2}}$ and substitute $z = \frac{x-m}{s}$:

$$\begin{aligned}
\tfrac{1}{\sqrt{2\pi s^2}} \int_l^u xe^{-\frac{(x-m)^2}{2s^2}} \mathrm{d}x &= \tfrac{s}{\sqrt{2\pi}} \int_{\frac{l-m}{s}}^{\frac{u-m}{s}} ze^{-\frac{z^2}{2}} \mathrm{d}z + \tfrac{m}{\sqrt{2\pi s^2}} \int_{\frac{l-m}{s}}^{\frac{u-m}{s}} e^{-\frac{z^2}{2}} \mathrm{d}z \\
&= m[\Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s})] - s[\phi(\tfrac{u-m}{s}) - \phi(\tfrac{l-m}{s})].
\end{aligned} \tag{21}$$

Putting it all together we have the expression for the integral I:

$$\begin{aligned}
I &= -\tfrac{1}{2}\log[2\pi\sigma^2][\Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s})] - \tfrac{s^2}{2\sigma^2}[\Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s}) - \tfrac{u-m}{s}\phi(\tfrac{u-m}{s}) + \tfrac{l-m}{s}\phi(\tfrac{l-m}{s})] \\
&\quad - \tfrac{2(m-\mu)}{2\sigma^2}\{m[\Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s})] - s[\phi(\tfrac{u-m}{s}) - \phi(\tfrac{l-m}{s})]\} - \tfrac{\mu^2-m^2}{2\sigma^2}[\Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s})] \\
&= \{\tfrac{m^2-\mu^2-2m(m-\mu)-s^2}{2\sigma^2} - \tfrac{1}{2}\log[2\pi\sigma^2]\}[\Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s})] + \tfrac{(u+m-2\mu)s}{2\sigma^2}\phi(\tfrac{u-m}{s}) - \tfrac{(l+m-2\mu)s}{2\sigma^2}\phi(\tfrac{l-m}{s}).
\end{aligned}$$

The formula for the cross entropy is:

$$
\begin{aligned}
&-\mathbb{E}_{\phi(m,s)} \log p(\mu, \sigma, l, u) \\
&\qquad = -\phi(\tfrac{l-m}{s}) \log[\Phi(\tfrac{l-\mu}{\sigma})] - I - \phi(\tfrac{u-m}{s}) \log[\Phi(\tfrac{\mu-u}{\sigma})] \\
&\qquad = \{\tfrac{1}{2}\log[2\pi\sigma^2] + \tfrac{(\mu-m)^2+s^2}{2\sigma^2}\}[\Phi(\tfrac{u-m}{s}) - \Phi(\tfrac{l-m}{s})] \\
&\qquad\quad - \{\log\Phi(\tfrac{\mu-u}{\sigma}) + \tfrac{(u+m-2\mu)s}{2\sigma^2}\}\phi(\tfrac{u-m}{s}) \\
&\qquad\quad - \{\log\Phi(\tfrac{l-\mu}{\sigma}) - \tfrac{(l+m-2\mu)s}{2\sigma^2}\}\phi(\tfrac{l-m}{s})
\end{aligned} \tag{22}
$$

D. Approximate Inference for Tobit Gaussian Process Regressors

Gaussian processes (GPs) are a popular choice of a prior for Bayesian non-parametric regression. Gaussian process is defined as a stochastic process indexed by a set $\mathcal{X} \in \mathbb{R}^d$: such that $\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$. Every finite combination of random variables of the process has a joint Gaussian distribution. Formally, GPs are distributions over functions. A GP is fully specified by its mean function μ and covariance structure over some finite set $\boldsymbol{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$, GP is uniquely defined by $p(\boldsymbol{f}|\mathcal{D}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{K})$ with latent vector $\boldsymbol{f} = (f(\boldsymbol{x_1}), \ldots, f(\boldsymbol{x_n}))$, mean vector $\boldsymbol{\mu} = (\mu(\boldsymbol{x_1}), \ldots, \mu(\boldsymbol{x_n}))$ and covariance matrix $\boldsymbol{K} = k(\boldsymbol{x_i}, \boldsymbol{x_j})_{i,j=1,\ldots n}$.

We denote the observed data as $\mathcal{D} = \{(\boldsymbol{x_i}, y_i), i = 1 \ldots n\}$, where $\boldsymbol{y} = (y_1, \ldots y_n)$ is the vector of labels. To simplify notation, we also denote by $\boldsymbol{\theta}$ the set of the Gaussian process parameters in general, e.g. $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{K})$.

D.1. Gaussian Process Model for Regression

To simplify the notation, we denote $f_i = f(\boldsymbol{x_i}), i = 1 \ldots n$. The standard GP regression task is focused on estimating the latent function $\boldsymbol{f}$ out of noisy observations $\boldsymbol{y} \in \mathbb{R}^n$: $y_i = f_i + \varepsilon_i, i = 1, \ldots, n$ and where $\varepsilon_i \sim \mathcal{N}(\varepsilon|0, \sigma_y^2)$.

To estimate the parameters, we can write the likelihood function, which is Gaussian due to the assumption on the noise term:

$$
p(\boldsymbol{y}|\boldsymbol{f}, \sigma_y) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2 \boldsymbol{I}_{nn}), \tag{23}
$$

where $\boldsymbol{I}_{nn}$ denotes $n \times n$ identity matrix.

To introduce Bayesian inference we assume zero mean prior GP with kernel function $k(\boldsymbol{x_i}, \boldsymbol{x_j})$: $f \sim GP(0, k(\boldsymbol{x_i}, \boldsymbol{x_j}))$. So the joint distribution of the latent function values corresponding to any set $\boldsymbol{X}$ of input data is a multivariate Gaussian:

$$
p(\boldsymbol{f}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{nn}), \tag{24}
$$

where $\boldsymbol{\theta}$ denotes the parameters of the kernel function and the noise parameter σ_y. Using Bayes rule we obtain for of posterior:

$$
p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{f})}{p(\mathcal{D}|\boldsymbol{\theta})} = \frac{\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{nn})}{p(\mathcal{D}|\boldsymbol{\theta})} \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2 \boldsymbol{I}_{nn}). \tag{25}
$$

The marginal likelihood can be evaluated as:

$$
p(\mathcal{D}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{f} = \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2 \boldsymbol{I}_{nn}) \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{nn}) d\boldsymbol{f}. \tag{26}
$$

Because we marginalise out from the two multivariate normals, the resulting distribution is also multivariate normal: $p(\boldsymbol{y}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{K}_{nn} + \sigma_y^2 \boldsymbol{I}_{nn})$. To estimate the parameters $\boldsymbol{\theta}$ of the marginal likelihood we can use gradient based optimisation to maximise the marginal log-likelihood of the model.

D.2. Gaussian Process for Binary Classification

While the inference for regression task has many analytic advantages, even the binary classification loses analytical tractability. For the binary classification task: y_i takes one of the two values $\{-1, 1\}$ only. The probability of the positive outcome $p(y_i = 1|\boldsymbol{x}) = \tilde{\sigma}(f(x))$, where $\tilde{\sigma}(\cdot)$ denotes the sigmoid transformation. For example, lets consider a probit model: $p(y_i = 1|\boldsymbol{x}) = \Phi(f(\boldsymbol{x}))$, where $\Phi(\cdot)$ is the CDF of standard normal distribution. The class labels are Bernoulli variables

with parameter $\Phi(f(x))$. The likelihood for the binary classification case then corresponds to the product distribution:

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{n} p(y_i|f_i) = \prod_{i=1}^{n} \Phi(y_i f_i) \tag{27}$$

Placing the GP prior over the latent function $\boldsymbol{f}$, as in (28) leads to posterior:

$$p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})}{p(\mathcal{D}|\boldsymbol{\theta})} \prod_{i=1}^{n} p(y_i|f_i) = \frac{\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{nn})}{p(\mathcal{D}|\boldsymbol{\theta})} \prod_{i=1}^{n} \Phi(y_i f_i) \tag{28}$$

Following similar methodology as for the regression task, we would like to evaluate the marginal likelihood:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{f} = \int \prod_{i=1}^{n} \Phi(y_i f_i) \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{nn}) d\boldsymbol{f}, \tag{29}$$

however due to the Bernoulli likelihood, the marginal likelihood is analytically intractable.

D.3. Gaussian Process Censored Regression

The censored regression problem corresponds to the case where the true target value $\eta \in \mathbb{R}$ are partially unobservable and instead we observe the censored version of the target y. For the censoring lower and upper thresholds: $l < u \in \mathbb{R}$, we describe the censoring process is described as:

$$y = \begin{cases} l, & \text{if } \eta \leq l, \\ \eta, & \text{if } l < \eta < u, \\ u, & \text{if } \eta \geq u, \end{cases} \tag{30}$$

i.e. the values inside the censoring range are shifted to the boundaries of the range and assigned a single value.

Consider the standard GP regression from D.1, where we estimate the latent function $\boldsymbol{f}$ out of noisy observations $\boldsymbol{\eta} \in \mathbb{R}^n$: $\eta_i = f_i + \varepsilon_i, i = 1, \ldots, n$ and where $\varepsilon_i \sim \mathcal{N}(\varepsilon|0, \sigma_y^2)$. Because the latent noisy targets are censored, our problem reformulates as: $y_i = f_i + \epsilon_i, i = 1, \ldots, n$, where $\epsilon_i \sim \mathcal{N}_c(\varepsilon|0, \sigma_y^2, l, u)$ comes from censored normal distribution with the censoring thresholds $l < u$. We can see this problem also as a combination of two probit models placed on the lower threshold $p(y_i^l = 1|\boldsymbol{x}) = \Phi(f(\boldsymbol{x})|l, \sigma_y^2)$, upper threshold $p(y_i^u = 1|\boldsymbol{x}) = 1 - \Phi(f(\boldsymbol{x})|u, \sigma_y^2)$, while having a normal regression model $p(\boldsymbol{y}|\boldsymbol{f}, \sigma_y) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2 \boldsymbol{I}_{nn})$ otherwise.

From this initial motivation, we can write the likelihood of the censored model as:

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{y_i \leq l} \Phi(l|f_i, \sigma_y^2) \prod_{l < y_i < u} \mathcal{N}(y_i|f_i, \sigma_y^2) \prod_{y_i \geq u} [1 - \Phi(u|f_i, \sigma_y^2)]. \tag{31}$$

We place the GP prior over the latent function $\boldsymbol{f}$ as in (24) and obtain the posterior distribution over these latent parameters $\boldsymbol{f}$:

$$\begin{aligned} p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\theta}) &= \frac{p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})}{p(\mathcal{D}|\boldsymbol{\theta})} \prod_{i=1}^{n} p(y_i|f_i) \\ &= \frac{\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{nn})}{p(\mathcal{D}|\boldsymbol{\theta})} \prod_{y_i \leq l} \Phi(l|f_i, \sigma_y^2) \prod_{l < y_i < u} \mathcal{N}(y_i|f_i, \sigma_y^2) \prod_{y_i \geq u} [1 - \Phi(u|f_i, \sigma_y^2)]. \end{aligned} \tag{32}$$

The marginal likelihood is analytically intractable due to the atoms on the censoring boundaries. Also note, that $\Phi(x|m, s^2) = 1 - \Phi(m|x, s^2)$. We will use this relationship to split the posterior into the truncated Gaussian part and the analytically

intractable mixture.

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}) &= \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta})d\boldsymbol{f} \\
&= \int \prod_{y_i \le l} \Phi(l|f_i,\sigma_y^2) \prod_{l<y_i<u} \mathcal{N}(y_i|f_i,\sigma_y^2) \prod_{y_i \ge u} [1-\Phi(u|f_i,\sigma_y^2)]\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K}_{nn})d\boldsymbol{f} \\
&= \int \mathcal{N}(\boldsymbol{y}\delta_{\boldsymbol{y}\in(l,u)}|\boldsymbol{f},\sigma_y^2)\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K}_{nn})d\boldsymbol{f} + \int \prod_{y_i \le l}[1-\Phi(f_i|l,\sigma_y^2)\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K}_{nn})d\boldsymbol{f} + \\
&+ \int \prod_{y_i \ge u} \Phi(f_i|u,\sigma_y^2)\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K}_{nn})d\boldsymbol{f} \\
&= \mathcal{N}(\boldsymbol{y}\delta_{\boldsymbol{y}\in(l,u)}|\boldsymbol{0},\boldsymbol{K}_{nn}+\sigma_y^2\boldsymbol{I}_{nn}) + \int \prod_{y_i \le l}[1-\Phi(f_i|l,\sigma_y^2)\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K}_{nn})d\boldsymbol{f} + \\
&+ \int \prod_{y_i \ge u} \Phi(f_i|u,\sigma_y^2)\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K}_{nn})d\boldsymbol{f}.
\end{aligned}
\tag{33}
$$

D.3.1. Variational Inference for Censored Regressor

Let us assume the Gaussian approximation of the posterior distribution:

$$
p(\boldsymbol{f}|\mathcal{D},\boldsymbol{\theta}) \approx q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}}) = \mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}},\boldsymbol{A}),
\tag{34}
$$

The goal is to minimize the KL divergence between the Gaussian approximation of the posterior and the posterior itself. In the following we provide the derivation of the evidence lower bound (ELBO) with the censored normal likelihood $p(\boldsymbol{y}|\boldsymbol{f})$ and Gaussian prior $p(\boldsymbol{f})$:

$$
\begin{aligned}
&KL\left[q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})||p(\boldsymbol{f}|\mathcal{D},\boldsymbol{\theta})\right] \\
&\quad= \int q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})\left[\ln q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}}) - \ln p(\boldsymbol{f}|\mathcal{D},\boldsymbol{\theta})\right]d\boldsymbol{f} \\
&\quad= \int q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})\left[\ln q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}}) - \ln p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta}) - \ln p(\boldsymbol{y}|\boldsymbol{f}) + \ln p(\mathcal{D}|\boldsymbol{\theta})\right]d\boldsymbol{f} \\
&\quad= KL\left[q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})||p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta})\right] - \int q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})\ln p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{f} + \ln p(\mathcal{D}|\boldsymbol{\theta}).
\end{aligned}
\tag{35}
$$

Because left handside term $KL\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\mathcal{D},\boldsymbol{\theta})\right]$ is non-negative and we aim to minimize this distance, we write the ELBO as:

$$
\begin{aligned}
\ln p(\mathcal{D}|\boldsymbol{\theta}) &\ge \mathbb{E}_{q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - KL\left[q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})||p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta})\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{q(f_i|\mathcal{D},\hat{\boldsymbol{\theta}})} \ln p(y_i|f_i) - KL\left[q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})||p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta})\right].
\end{aligned}
\tag{36}
$$

The first term comprises the (negative) cross entropies of the approximate posterior with respect to the likelihood. The cross entropy can be split into two terms: a factorised likelihood and a marginalisation term of the remaining variables. We can

then apply the derived crossed entropy between Gaussian and censored distribution. Let us derive the analytic form:

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}_{q(f_i|\mathcal{D},\hat{\boldsymbol{\theta}})}\log p(y_i|f_i)\\
&\quad=-\sum_{i=1}^{n}\Big(\log[\sqrt{2\pi}\sigma_y]+\tfrac{(f_i-\hat{f}_i)^2+a_{ii}^2}{2\sigma_y^2}\Big)\Big[\Phi\Big(\tfrac{u-\hat{f}_i}{a_{ii}}\Big)-\Phi\Big(\tfrac{l-\hat{f}_i}{a_{ii}}\Big)\Big]\\
&\qquad+\sum_{i=1}^{n}\Big[\log\Phi\Big(\tfrac{f_i-u}{\sigma_y}\Big)+\tfrac{(u+\hat{f}_i-2f_i)a_{ii}}{2\sigma_y^2}\Big]\phi\Big(\tfrac{u-\hat{f}_i}{a_{ii}}\Big)\\
&\qquad+\sum_{i=1}^{n}\Big[\log\Phi\Big(\tfrac{l-f_i}{\sigma_y}\Big)-\tfrac{(l+\hat{f}_i-2f_i)a_{ii}}{2\sigma_y^2}\Big]\phi\Big(\tfrac{l-\hat{f}_i}{a_{ii}}\Big).
\end{aligned}\tag{37}$$

The second term is the KL divergence between two multivariate Gaussian distributions. It can be written in matrix calculus as follows:

$$\begin{aligned}
KL\Big[q(\boldsymbol{f}|\mathcal{D},\hat{\boldsymbol{\theta}})||p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta})\Big]&=KL\Big[\mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}},\boldsymbol{A})||\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K_{nn}})\Big]\\
&=\mathbb{E}_{\mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}},\boldsymbol{A})}\Big[\ln\mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}},\boldsymbol{A})-\ln\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K_{nn}})\Big]\\
&=\frac{1}{2}\mathbb{E}_{\mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}},\boldsymbol{A})}\Big[-\ln|\boldsymbol{A}|-(\boldsymbol{f}-\hat{\boldsymbol{f}})^T\boldsymbol{A}^{-1}(\boldsymbol{f}-\hat{\boldsymbol{f}})+\ln|\boldsymbol{K_{nn}}|+\boldsymbol{f}^T\boldsymbol{K_{nn}^{-1}}\boldsymbol{f}\Big]\\
&=\frac{1}{2}\ln\frac{|\boldsymbol{A}|}{|\boldsymbol{K_{nn}}|}+\frac{1}{2}\mathbb{E}_{\mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}},\boldsymbol{A})}\Big[-\mathrm{tr}[\boldsymbol{A}^{-1}(\boldsymbol{f}-\hat{\boldsymbol{f}})(\boldsymbol{f}-\hat{\boldsymbol{f}})^T]+\mathrm{tr}[\boldsymbol{K_{nn}^{-1}}\boldsymbol{f}\boldsymbol{f}^T]\Big]\\
&=\frac{1}{2}\ln\frac{|\boldsymbol{A}|}{|\boldsymbol{K_{nn}}|}-\frac{1}{2}\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^{-1})+\int\mathrm{tr}[\boldsymbol{K_{nn}^{-1}}\boldsymbol{f}\boldsymbol{f}^T]\mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}},\boldsymbol{A})d\boldsymbol{f}\\
&=\frac{1}{2}\Big[\ln|\boldsymbol{A}|-\ln|\boldsymbol{K_{nn}}|-n+\mathrm{tr}(\boldsymbol{K_{nn}^{-1}}\boldsymbol{A})+\hat{\boldsymbol{f}}^T\boldsymbol{K_{nn}^{-1}}\hat{\boldsymbol{f}}\Big].
\end{aligned}\tag{38}$$

We arrive to the final version of the ELBO:

$$\ln p(\mathcal{D}|\boldsymbol{\theta})\geq-\sum_{i=1}^{n}\Big(\log[\sqrt{2\pi}\sigma_y]+\tfrac{(f_i-\hat{f}_i)^2+a_{ii}^2}{2\sigma_y^2}\Big)\Big[\Phi\Big(\tfrac{u-\hat{f}_i}{a_{ii}}\Big)-\Phi\Big(\tfrac{l-\hat{f}_i}{a_{ii}}\Big)\Big]\tag{39}$$

$$+\sum_{i=1}^{n}\Big[\log\Phi\Big(\tfrac{f_i-u}{\sigma_y}\Big)+\tfrac{(u+\hat{f}_i-2f_i)a_{ii}}{2\sigma_y^2}\Big]\phi\Big(\tfrac{u-\hat{f}_i}{a_{ii}}\Big)\tag{40}$$

$$+\sum_{i=1}^{n}\Big[\log\Phi\Big(\tfrac{l-f_i}{\sigma_y}\Big)-\tfrac{(l+\hat{f}_i-2f_i)a_{ii}}{2\sigma_y^2}\Big]\phi\Big(\tfrac{l-\hat{f}_i}{a_{ii}}\Big)\tag{41}$$

$$-\frac{1}{2}\Big[\ln|\boldsymbol{A}|-\ln|\boldsymbol{K_{nn}}|-n+\mathrm{tr}(\boldsymbol{K_{nn}^{-1}}\boldsymbol{A})+\hat{\boldsymbol{f}}^T\boldsymbol{K_{nn}^{-1}}\hat{\boldsymbol{f}}\Big].\tag{42}$$