
COBALT: Censored Optimization and Bayesian Active Learning Techniques

Andrea Karlová¹

Rishabh Kabra^{1,2}

Daniel Augusto de Souza¹

Brooks Paige¹

¹University College London, London, UK

²DeepMind, London, UK,

Abstract

We target Bayesian Active Learning (AL) and Optimization (BO) for censored data regimes. While the Tobit likelihood accurately models such clipped observations, its mixed continuous-discrete nature impedes the analytical evaluation of information-theoretic acquisition functions. To address this, we investigate Censored Optimization via Bayesian Active Learning Techniques (COBALT). We establish rigorous theoretical guarantees for this framework, proving posterior consistency and the asymptotic normality of the MAP estimator under greedy maximization. Central to our framework is the derivation of a closed-form entropy for the Censored Normal distribution, enabling an analytical BALD (Bayesian Active Learning by Disagreement) score compatible with any Gaussian posterior approximations. We further underpin this method by deriving a numerically stable Evidence Lower Bound (ELBO) for censored atoms, utilizing robust approximations of the log-normal cumulative density. Empirical evaluations using our open-source implementation demonstrate COBALT’s best accuracy–compute trade-off among censored-likelihood methods in learning GP posteriors and effectiveness on a variety of benchmarks. Code: <https://github.com/AndreaKarlova/cobalt>.

1 INTRODUCTION

Real-world measurements often come from bounded-scale instruments, resulting in data being censored not as a statistical anomaly, but as a structural property. Consider an example of binding affinity measurement in drug discovery [Svensson et al., 2025]. High-throughput assays quantify potent ligands but “clip” non-binders at a limit of detection l , yielding observations where the true affinity y^* is unknown beyond the fact that $y^* \geq l$. Chaining a binary classifier

with a regressor trained only on the binders introduces systematic bias by effectively truncating the data distribution. This approach ignores the constraints from non-binders and degrades uncertainty quantification near the decision boundary. Using Tobit likelihood [Tobin, 1958] allows to model the mixed continuous (regression) and discrete (classifier) nature of the data. Gaussian Processes (GPs) are well suited for the uncertainty quantification, e.g. to enable researchers make risk-aware decisions in low-data regimes in highly noisy, partially observable problems. GPs well serve as effective surrogates for both AL and BO, with practical BO applications e.g. in hyper-parameters optimization with early stopping scheduler [Hutter et al., 2013].

However, using the Tobit likelihood within the adaptive framework presents a significant theoretical challenge. While standard GPs provide closed-form expressions for entropy and mutual information, Censored GP traditionally do not. Although the Tobit likelihood corrects the systematic bias, it makes the posterior analytically intractable, which prevents the direct computation of the exact information-theoretic quantities needed for AL and BO. We address this gap by providing a solid theoretical basis for information-theoretic design under censoring and present analytic approximations for relevant quantities.

Contributions.

1. **Theoretical Guarantees for Censored AL.** We identify the regularity and identifiability conditions under which the asymptotic guarantees of information-theoretic design [Paninski, 2005] transfer to GPs with a Tobit likelihood, and verify them for the censored-normal model (Lemmas 4.1–4.2). Under these conditions the greedy info-max design is consistent and asymptotically D-optimal: the volume of its confidence ellipsoid is no greater than that of random sampling, and in general smaller.
2. **Closed-Form BALD Approximation and Censored Predictive Entropy Search.** We derive the exact closed-form entropy for censored-normal distribution (Lemma 3.1) and use it to obtain Censored BALD score (Lemma

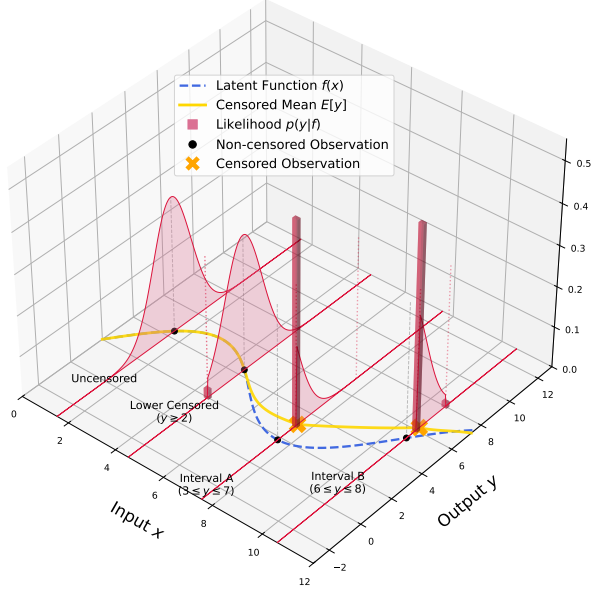


Figure 1: **Censored Likelihoods with Varying Bounds.** The blue dashed trace depicts the latent function $f(x)$, the solid yellow line represents the censored (observed) mean of $y(x)$. The censoring applied to the observations is captured by the orange cross. The crimson vertical slices illustrate the observation likelihood $p(y_i | f_i)$ under different censoring regimes: (1) **Uncensored:** Gaussian likelihood. (2) **Lower Censored:** The distribution is truncated below $y = 2$, with the cut-off tail probability concentrated into a discrete mass (red vertical bar). (3) & (4) **Interval Censored:** The distributions are confined to $[3, 7]$ and $[6, 8]$ respectively, with probability masses accumulating at both boundaries.

4.3) under Gaussian approximation of the posterior. For optimization we derive a Censored PES acquisition whose predictive-entropy term is analytic via Lemma 3.1.

3. **Numerically Stable Analytic ELBO.** We derive a numerically stable analytic ELBO for the Tobit likelihood that avoids the tail instability of inverse-Mills-ratio evaluations, using a log-CDF lower bound that remains stable in the deep tails where naive computation underflows. We also support alternative Gaussian approximations such as Laplace approximation (LA) and Expectation Propagation (EP).

2 BACKGROUND

Adaptive Bayesian framework involves critical design choices: (i) a prior distribution $p(f)$ over the latent function space, (ii) a likelihood $p(y_i | f)$ capturing the data generation process, (iii) an inference method for the posterior distribution $p(f | y_1, \dots, y_N) \propto \prod_i p(y_i | f)p(f)$ given observed data, and (iv) a data acquisition strategy to aid further learning. The choice of likelihood extends beyond Bayesian

inference; it defines the loss landscape for any regression task. Using a standard Gaussian likelihood (equivalent to minimizing Mean Squared Error) on censored data introduces systematic bias regardless of the model architecture. Consequently, our primary focus is the Tobit likelihood specification, which we adapt to the general case where censoring bounds vary per observation. Figure 1 visually demonstrates this challenge: the likelihood must dynamically transition between standard Gaussian densities and truncated forms with discrete mass accumulations depending on the local bounds at input x .

2.1 A LIKELIHOOD FOR CENSORED DATA

In a censored regression setting, we observe a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i, c_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ are the inputs and $y_i \in \mathbb{R}$ are the targets. Unlike standard Gaussian regression, where observations are assumed to be precise measurements of a latent function $f(\mathbf{x}_i)$ corrupted by noise, the data here may be *censored*. This is captured by an indicator variable $c_i \in \{-1, 0, 1\}$ for each datapoint. The case $c_i = 0$ denotes a direct observation, while $c_i = -1$ and $c_i = 1$ indicate left-censoring (a lower bound) and right-censoring (an upper bound), respectively. The observation process generalizes the Tobit model [Tobin, 1958]. While this approach has been extensively studied in econometrics [Amemiya, 1984, Robin, 2010] and survival analysis [Klein et al., 2003], it has also been extended to handle various censoring mechanisms in machine learning [Pearce et al., 2022, Friederichs and Hense, 2007, Moradian et al., 2017]. Specifically, we allow for varying bounds specific to each data point:

$$y_i = \begin{cases} l_i & \text{if } c_i = -1 \quad \text{for } f(\mathbf{x}_i) + \varepsilon \leq y_i \\ f(\mathbf{x}_i) + \varepsilon & \text{if } c_i = 0 \quad \text{for } f(\mathbf{x}_i) + \varepsilon = y_i \\ u_i & \text{if } c_i = 1 \quad \text{for } f(\mathbf{x}_i) + \varepsilon \geq y_i \end{cases} \quad (1)$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The resulting likelihood is not absolutely continuous with respect to the Lebesgue measure, as it combines a continuous density for uncensored points with discrete probability masses for censored intervals. Consequently, the likelihood function $p(y_i | f_i, c_i)$, where we denote $f_i = f(\mathbf{x}_i)$ for brevity, is defined as:

$$p(y_i | f_i, c_i) = \begin{cases} \Phi\left(\frac{y_i - f_i}{\sigma}\right) & \text{if } c_i = -1 \\ \frac{1}{\sigma} \phi\left(\frac{y_i - f_i}{\sigma}\right) & \text{if } c_i = 0 \\ 1 - \Phi\left(\frac{y_i - f_i}{\sigma}\right) & \text{if } c_i = 1 \end{cases}, \quad (2)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the PDF and CDF of the standard normal distribution, respectively.

2.2 GPS AND POSTERIOR APPROXIMATIONS

We encode our prior beliefs over the function space using a Gaussian Process, a distribution built on the as-

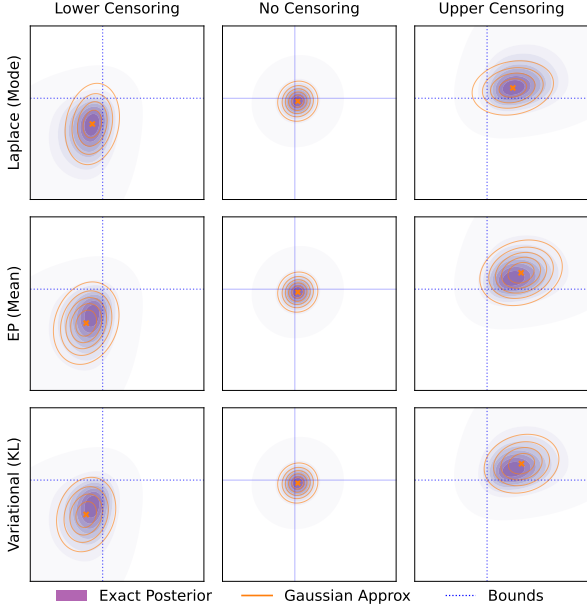


Figure 2: **Gaussian Approximations across Censoring Regimes.** We show the exact posterior (violet fill) over latent function values f_1 and f_2 for three different observation models (columns) and inference methods (rows). The likelihood noise $\sigma_n = 0.4$. The left column depicts *Lower Censoring* ($f_i \leq y_i$), the middle column shows standard *Gaussian noise* (no censoring), and the right column illustrates *Upper Censoring* ($f_i \geq y_i$). The rows show Gaussian approximations (orange contours) obtained via LA (centered at the mode), EP (matching moments), and VI (minimizing KL divergence). The blue dotted lines indicate the censoring boundaries placed into the data locations.

sumption that any two values $f(x)$ and $f(x')$ are correlated Gaussian random variables [Williams and Rasmussen, 2006]. Given training and evaluation sets $\mathbf{X}$ and $\mathbf{X}^*$, we represent the output of the latent functions as $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)]$ and $\mathbf{f}^* = [f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_N^*)]$ with a joint distribution parametrized by $[\mu_f]_i = \mu(x_i)$ and $[\mathbf{K}_f]_{i,j} = k(x_i, x_j)$. GPs are favored because their conjugacy with the Gaussian likelihood permits exact, closed-form inference. This yields analytical expressions for the posterior $p(\mathbf{f}|\mathbf{y})$, predictive posterior $p(\mathbf{f}^*|\mathbf{y})$, and model evidence $p(\mathbf{y}) = \mathbb{E}_{p(\mathbf{f})} [p(\mathbf{y} | \mathbf{f})]$, enabling gradient-based selection of hyperparameters by maximizing the evidence.

For non-Gaussian likelihoods, such as the censored likelihood described above, exact inference is intractable; the posterior distribution $p(\mathbf{f} | \mathbf{y})$ cannot be computed analytically. Typically, we approximate the intractable posterior with a tractable Gaussian distribution $q(\mathbf{f}) = \mathcal{N}(\mathbf{f} | \mathbf{m}, \mathbf{V})$. Crucially, the censored likelihood is log-concave with respect to the latent function f . This geometric property guarantees that the posterior distribution is unimodal and justifies approximating it with a Gaussian [Nickisch and Rasmussen,

2008]. While several strategies exist to determine the parameters $\mathbf{m}$ and $\mathbf{V}$, they represent different trade-offs between computational stability and predictive accuracy:

Laplace Approximation (LA) estimates the posterior $p(\mathbf{f}|\mathbf{y})$ by a Gaussian $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{V})$ centered at the Maximum A Posteriori (MAP) estimate $\mathbf{m} = \arg \max_{\mathbf{f}} \log p(\mathbf{f}|\mathbf{y})$. The covariance is approximated via the second-order Taylor approximation around the mode $\mathbf{m}$ as the inverse of the negative Hessian $\mathbf{V} = (-\nabla^2 \log p(\mathbf{f}|\mathbf{y})|_{\mathbf{f}=\mathbf{m}})^{-1} = (\mathbf{K}^{-1} + \mathbf{W})^{-1}$, where $\mathbf{K}$ is the kernel matrix and $\mathbf{W}$ is the diagonal matrix of the likelihood’s second derivatives. While computationally effective for log-concave likelihoods, LA relies solely on local curvature at the peak. It tends to underestimate the posterior variance [Hartmann and Vanhatalo, 2019], particularly in high-dimensional spaces where the probability mass is not concentrated at the mode [Opper and Archambeau, 2009].

Expectation Propagation (EP) approximates the global posterior by replacing each likelihood factor with a local Gaussian site, $t_i(f_i) \propto \mathcal{N}(\tilde{\mu}_i, \tilde{\sigma}_i^2)$, whose parameters are iteratively updated to match the zeroth, first, and second marginal moments of the “tilted” distribution—the product of the true likelihood and the cavity distribution [Minka, 2001]. While this local moment-matching strategy frequently yields predictive accuracy comparable to MCMC [Nickisch and Rasmussen, 2008, Gammelli et al., 2020], it introduces significant instability in censored regimes. The requisite integrals for the Tobit likelihood involve the hazard function (inverse Mills ratio), $\frac{\mathcal{N}(z)}{\Phi(z)}$, which is numerically volatile in the tails; consequently, EP implementations require specialized log-space expansions and heavy damping to prevent the site approximations oscillating between “observed” and “clipped” states [Vanhatalo et al., 2009].

Variational Inference (VI) approximates the true posterior by maximizing the Evidence Lower Bound (ELBO), $\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{f})} [\log p(\mathbf{y}|\mathbf{f})] - \text{KL}(q(\mathbf{f})||p(\mathbf{f}))$. The ELBO provides a rigorous, differentiable lower bound on the model evidence. Unlike EP, which relies on iterative moment matching and can oscillate, KL minimization transforms inference into a well-posed optimization problem guaranteed to converge [Titsias, 2009, Hensman et al., 2013]. The “zero-forcing” property of the reverse KL typically yields more compact variance estimates than EP’s mass-covering approximation [Minka, 2005, Nickisch and Rasmussen, 2008]. This enables principled model selection and hyperparameter learning better suited for adaptive systems, where the numerical instability of EP is a liability.

2.3 RELATED WORK

Initial work by Ertin [2007] on Censored GP Regression considered one-sided constant censoring bound and Laplacian approximate inference. This was later extended to two-

sided constant censoring bound using Expectation Propagation (EP) Groot and Lucas [2012], while Gammelli et al. [2020] explored random upper censored bounds for shared mobility and demand modeling. Further, Gammelli et al. [2022] explored highly generalized frameworks to handle heteroscedastic noise and multi-output regression. However, their method relied on Monte Carlo sampling which can be slow compared to analytic approximations. Basson et al. [2023] developed a closed-form variational lower bound using sparse variational GPs for scalability, yet their solution introduces an artificial parameter, ξ , which critically impacts fit quality. Motivated by the high variance and lack of control in former approaches and the parameter-dependent fits of the latter, we leverage the KL-minimization framework and implemented a numerically robust analytical ELBO, providing a closed-form approximation of the Evidence Lower Bound (ELBO) for the censored Gaussian likelihood. The drawbacks however are the lower speed and tendency to over-confidence [Nickisch and Rasmussen, 2008].

Beyond censoring-specific literature, recent advances in GP inference with non-conjugate likelihoods have re-examined the trade-offs between different approximation schemes. Li et al. [2023] demonstrated that while VI provides a stable objective for posterior approximation, jointly optimizing hyperparameters via the Evidence Lower Bound (ELBO) can introduce bias, often underperforming compared to EP for model selection. They proposed a hybrid training procedure that decouples inference from hyperparameter learning to mitigate this issue. However, such hybrid approaches often require complex implementation and iterative updates that may not be suitable for all censored regimes.

A parallel line of research has begun to address decision-making in censored regimes. While standard Bayesian Active Learning by Disagreement (BALD) and Predictive Entropy Search (PES) are well-established for Gaussian targets [Houlsby et al., 2011, Hernández-Lobato et al., 2014], their application to clipped data remains limited. Recently, [Hüttel et al., 2025] have introduced variants such as C-BALD, which relies on Bayesian Neural Networks (BNN), MC Dropout for uncertainty quantification and addresses scenarios where the censoring mechanism itself is random. BNNs are designed for high-data regimes, whereas we target low-data regimes where GP-based surrogates are good modeling practice and makes application relevant to BO.

3 METHOD

Standard frameworks for purely discrete or purely continuous distributions are not directly applicable to a censored probability measure. While methods like Nair et al. [2006] model this as a “mixed pair” to analyze entropy rates, we adopt a more rigorous measure-theoretic framework based on the Lebesgue decomposition theorem [Kallenberg, 2021]. Rather than treating the likelihood as a simple mixture, we

treat the continuous regions separately from the discrete point masses by defining the probability density via the Radon-Nikodym derivative with respect to a reference measure $\rho = \lambda + \delta_\ell + \delta_u$, where λ is the Lebesgue measure and δ represents the Dirac measures at the censoring bounds. This ensures the entropy is well-defined and integrates the continuous density with the discrete boundary masses.

For a censored normal likelihood with latent mean f , noise σ , and censoring bounds $[\ell, u]$, the likelihood $p(y | f, \sigma, \ell, u)$ is defined as:

$$p(y | f, \sigma, \ell, u) = \phi(y | f, \sigma^2) \mathbb{I}_{(\ell, u)}(y) + \Phi\left(\frac{\ell-f}{\sigma}\right) \delta_\ell(y) + \left(1 - \Phi\left(\frac{u-f}{\sigma}\right)\right) \delta_u(y) \quad (3)$$

where Φ is the standard normal CDF. Based on this measure-theoretic foundation, we derive the closed-form entropy for the Censored Normal distribution.

Lemma 3.1 (Entropy of the Censored Normal Distribution). *Let Y be a random variable following a Censored Normal distribution with parameters μ, σ and bounds ℓ, u . The entropy $H[Y]$ wrt to the mixed measure $\rho = \lambda + \delta_u + \delta_\ell$ is given by:*

$$H[Y] = \log(\sqrt{2\pi e}\sigma) \cdot \Delta\Phi - \frac{1}{2} [\zeta_u \phi(\zeta_u) - \zeta_\ell \phi(\zeta_\ell)] - \Phi(\zeta_\ell) \log \Phi(\zeta_\ell) - \Phi(-\zeta_u) \log \Phi(-\zeta_u) \quad (4)$$

where $\zeta_x = \frac{x-\mu}{\sigma}$ are the standardized bounds, $\Delta\Phi = \Phi(\zeta_u) - \Phi(\zeta_\ell)$ is the probability mass within the observation window, and ϕ is the standard normal PDF.

Proof. See Appendix B.2 for the full derivation using the Radon-Nikodym derivative. $\square$

The first term represents the entropy of the standard normal scaled by the mass of the observable interval. The second term accounts for the skew due to asymmetric censoring: this term vanishes if the bounds are symmetric around the mean. The final terms represent entropy contributions from the discrete probability masses accumulated at the tails.

Suppose we have a probabilistic model which marginally yields a latent Gaussian distribution over f , denoted $q(f | m, s) = \mathcal{N}(f | m, s^2)$. To construct an ELBO, we need to compute the expected log-likelihood wrt to the posterior approximation. For a single datapoint y_i , this expectation decomposes into a continuous term for uncensored observations and discrete mass terms for the censored boundaries:

$$\begin{aligned} \mathbb{E}_{q(f|m,s)} p(y | f, \sigma, \ell, u) = & \mathbb{E}_{q(f|m,s)} \log \phi(y | f, \sigma^2) \mathbb{I}_{(\ell, u)}(y) \\ & + \mathbb{E}_{q(f|m,s)} \log \Phi\left(\frac{\ell-f}{\sigma}\right) \delta_\ell(y) + \\ & + \mathbb{E}_{q(f|m,s)} \log \left(\Phi\left(\frac{f-u}{\sigma}\right)\right) \delta_u(y) \end{aligned} \quad (5)$$

The first term, for uncensored observations where $y \in (\ell, u)$, admits a straightforward closed-form result, while the expectations for the censored atoms involve terms of the form $\mathbb{E}_{\mathcal{N}(z|0,1)}[\log \Phi(a + bz)]$. Evaluating these is numerically challenging because $\log \Phi(z)$ underflow for moderately negative values. To address this, we use a lower bound derived from the Gaussian Q-function, $\log \Phi(z) > \log(-z) - \log(1 + z^2) - z^2/2 - \log(2\pi)/2$, for values $z < -5$ where the standard computation becomes unstable. We explored Gaussian-Hermite quadrature, Simpson’s rule, and the trapezoidal rule as integration tools. See the Supplementary for a detailed discussion on the robustness and implementation of these quadrature methods.

4 ACTIVE LEARNING WITH A CENSORED REGRESSOR

Our information-theoretic approach to active learning is grounded in the asymptotic theory established by Paninski [2005], which rigorously justifies the “greedy” maximization of mutual information. Paninski extends the Bernstein-von Mises theorem to the adaptive setting, proving that the posterior distribution asymptotically converges to a normal distribution centered on the true parameters. This strategy is shown to be asymptotically efficient, yielding parameter estimates with a variance strictly lower than or equal to that of non-adaptive (random) designs.

These guarantees rely on implicit regularity conditions, such as the compactness of the parameter space and the Lipschitz continuity of the likelihood functions. Without these constraints, the greedy approach can become inconsistent, particularly in unbounded or discontinuous settings. We formalize the assumptions under which the theoretical guarantees hold for Censored GPs in the following lemma:

Lemma 4.1 (Consistency of Information-Theoretic Design with Censored Likelihoods). *Let f be an element of a parameter space $\Theta \subseteq C(\mathcal{X})$ equipped with a supremum metric ρ , and let $f_0 \in \Theta$ be the true function. The posterior distribution $p_N(f|y)$, estimated on N points, converges in probability to f_0 as $N \rightarrow \infty$ under the information-maximization strategy if the following conditions hold:*

- (C1) **Compactness:** *The input domain $\mathcal{X}$ is compact and the parameter space Θ is closed, uniformly bounded and equicontinuous family of functions.*
- (C2) **Regularity:** *The censoring link function is uniformly Lipschitz continuous (i.e., has a bounded derivative).*
- (C3) **Prior Support:** *The prior p_0 assigns positive mass to any ϵ -neighborhood of f_0 .*
- (C4) **Identifiability:** *For all $f \neq f_0$, there exists an input distinguishing the functions:*

$$\sup_{x \in \mathcal{X}} D_{\text{KL}}(f_0; f | x) > 0.$$

Proof. See Appendix D.1. $\square$

The assumption Compactness assumption (C1) is standard in the BO literature [Shahriari et al., 2016]. For a GP prior this holds whenever the kernel is sufficiently regular that sample paths are almost surely Lipschitz on X [Srinivas et al., 2010, Shahriari et al., 2016].

Building on this consistency, we can further characterize the asymptotic behavior of the uncertainty estimates.

Lemma 4.2 (Asymptotic Normality and MAP Consistency of GP Projections). *Under the conditions of Lemma 4.1, consider the GP evaluated at a finite set of N query points, denoted by the vector $\mathbf{f}_N = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)]^\top$. Assuming the censored likelihood link function is twice continuously differentiable (C^2), the posterior of this finite projection $p(\mathbf{f}_N | \mathcal{D}_N)$ converges in total variation to a multivariate normal $\mathcal{N}(\mathbf{f}_{0,N}, \Sigma_N)$ centered on the true function values. Consequently, the MAP estimator $\hat{\mathbf{f}}_{\text{MAP}}$ is consistent almost surely and asymptotically distributed as $\mathcal{N}(\mathbf{f}_{0,N}, \Sigma_N)$. Furthermore, the determinant of the asymptotic covariance Σ_N under the information-maximization strategy satisfies:*

$$|\Sigma_{\text{info}}| \leq |\Sigma_{\text{iid}}|, \quad (6)$$

where Σ_{iid} is the covariance resulting from random sampling. This implies that the active learning strategy minimizes the volume of the confidence ellipsoid for the latent function values more efficiently than random sampling.

Proof. See Appendix D.2. $\square$

With this theoretical framework ensuring the validity of the method, our task is to efficiently reduce the space of possible hypotheses by selecting query points that maximize the reduction in entropy of the posterior distribution over the latent function. We select an input x that minimizes the entropy $H[f | \mathcal{D} \cup \{(x, y)\}]$, where the posterior $p(f | \mathcal{D} \cup \{(x, y)\})$ is the distribution resulting from observing the new label (x, y) . Following [Houlsby et al., 2011], this objective is formulated via the mutual information between the model parameters and the unknown outputs:

$$I(y; f | x, \mathcal{D}) = H[y | x, \mathcal{D}] - \mathbb{E}_{p(f|x,\mathcal{D})}[H(y | f, x)]. \quad (7)$$

Using a variational approximation of the posterior $q(f | x) = \mathcal{N}(f | m_x, s_x^2)$, we can explicitly quantify this mutual information, known as Bayesian Active Learning by Disagreement (BALD). For our Censored Normal likelihood with bounds $[l, u]$ and noise variance σ_y^2 , we derive the analytic score as follows:

Lemma 4.3 (Analytical Censored BALD and its Approximation). *For an input x , let the latent approximate posterior be $q(f | x) = \mathcal{N}(m_x, s_x^2)$, and let observations be censored to $[l, u]$ with homoscedastic noise variance σ_y^2 . Let*

$v_{xy} = \sqrt{s_x^2 + \sigma_y^2}$ denote the total predictive standard deviation. We define the standardized distances to the bounds as:

$$\alpha_{xy} = \frac{l-m_x}{v_{xy}}, \quad \beta_{xy} = \frac{u-m_x}{v_{xy}}.$$

The BALD score is given by:

$$\begin{aligned} \text{BALD}(x) = & \frac{1}{2} \log \left(1 + \frac{s_x^2}{\sigma_y^2} \right) [\Phi(\beta_{xy}) - \Phi(\alpha_{xy})] \\ & - \frac{s_x^2}{2v_{xy}^2} [\beta_{xy} \phi(\beta_{xy}) - \alpha_{xy} \phi(\alpha_{xy})] \\ & - \Phi(\alpha_{xy}) \log \Phi(\alpha_{xy}) - \Phi(-\beta_{xy}) \log \Phi(-\beta_{xy}) \\ & + \mathbb{E}_{z \sim \mathcal{N}(0,1)} [\Psi(z)], \end{aligned}$$

where the residual term Ψ captures the entropic contribution of the censored regions:

$$\begin{aligned} \Psi(z) = & \Phi \left(\frac{l-m_x+s_x z}{\sigma_y} \right) \log \Phi \left(\frac{l-m_x+s_x z}{\sigma_y} \right) \\ & + \Phi \left(\frac{m_x-u+s_x z}{\sigma_y} \right) \log \Phi \left(\frac{m_x-u+s_x z}{\sigma_y} \right). \end{aligned} \quad (8)$$

This integral term admits a closed-form Gaussian convolution approximation. With constants $A = 1/e$, $\mu_0 = \Phi^{-1}(1/e)$, and $\sigma_0 \approx 0.9794$:

$$\begin{aligned} \mathbb{E}_{z \sim \mathcal{N}(0,1)} [\Psi(z)] \approx & \frac{A \sigma_0 \sigma_y}{\sqrt{s_x^2 + \sigma_y^2 \sigma_0^2}} \exp \left(-\frac{(l-m_x-\sigma_y \mu_0)^2}{2(s_x^2 + \sigma_y^2 \sigma_0^2)} \right) \\ & - \frac{A \sigma_0 \sigma_y}{\sqrt{s_x^2 + \sigma_y^2 \sigma_0^2}} \exp \left(-\frac{(m_x-u-\sigma_y \mu_0)^2}{2(s_x^2 + \sigma_y^2 \sigma_0^2)} \right). \end{aligned} \quad (9)$$

Proof. See Appendix C.2 for the full derivation and numerical validation of the approximation. $\square$

The approximation error in (9) is the largest exactly where it matters the least for query selection. For very small model uncertainty s_x (e.g. $s_x \ll 0.1$), the error can reach up to .02 nats, but in this regime the true mutual information is itself bounded and close to zero, so such points are never competitive for acquisition. On the other hand, in the high-variance, unexplored regions ($s_x \gg 1$) that greedy BALD maximization targets, the approximation error is empirically below 10^{-3} nats. Since the true information gain in these uncertain regions outweighs both the error and the information gain available in well-explored, low-uncertainty regions, the difference between the approximate and exact scores is not expected to change the arg max in practice. Thus, the approximation error should not accumulate during the course of active learning, see Appendix C.2.

With $l = -\infty$ and $u = +\infty$, the censoring terms vanish, and the formula reduces to the standard BALD objective for

Normal Likelihood: $\frac{1}{2} \log \left(1 + \frac{s_x^2}{\sigma_y^2} \right)$. Conversely, for symmetric heavy censoring, the terms involving the probability masses at the bounds dominate, reducing to the expectation of the binary entropy of a Bernoulli variable.

5 PREDICTIVE ENTROPY SEARCH FOR CENSORED OPTIMIZATION

In the context of BO, our goal shifts from learning the global function parameters to identifying the global optimizer $\mathbf{x}_* = \arg \max_{\mathbf{x}} f(\mathbf{x})$. We employ Predictive Entropy Search (PES) [Hernández-Lobato et al., 2014], which selects the query point that leads to the largest reduction in the entropy of the posterior distribution over the global maximum. The acquisition function is defined as the mutual information between the observation y and optimum $\mathbf{x}_*$:

$$\alpha_{\text{PES}}(\mathbf{x}) = H[y | \mathbf{x}, \mathcal{D}] - \mathbb{E}_{p(\mathbf{x}|\mathcal{D})} H[y | \mathbf{x}, \mathbf{x}_*, \mathcal{D}]. \quad (10)$$

Using Gaussian approximation of the posterior $q(f | \mathbf{x}) = \mathcal{N}(f | m_x, s_x^2)$, we can explicitly quantify the components of this mutual information. For our Censored Normal likelihood with bounds $[l, u]$ and noise variance σ_y^2 , we can apply the analytic entropy expression derived in Lemma 3.1, we can evaluate the predictive entropy term as:

$$H[y | \mathbf{x}, \mathcal{D}] = H[\mathcal{CN}(m_x, s_x^2 + \sigma_y^2; l, u)]. \quad (11)$$

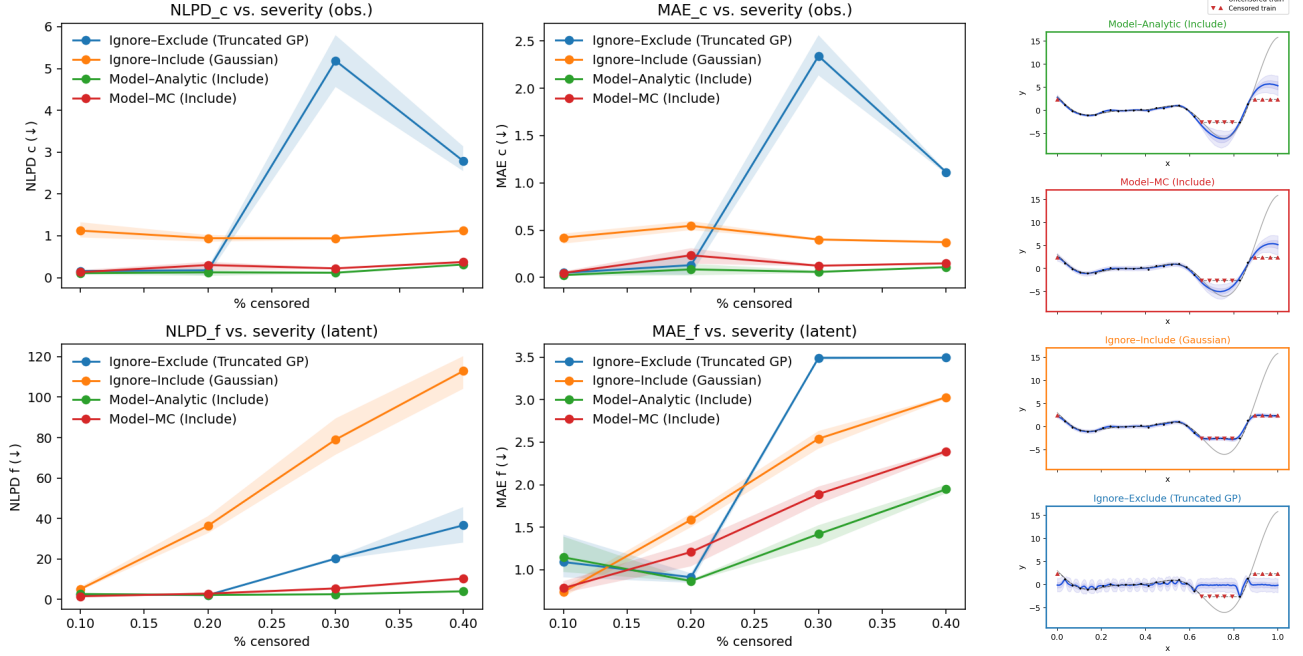
The second term requires evaluating the expected conditional entropy over the unknown global optimizer $\mathbf{x}_*$. Since this expectation is analytically intractable, we approximate it via Monte Carlo integration. Thompson Sampling [Thompson, 1933], we generate K samples of the global optimizer, $\{\mathbf{x}_*^{(k)}\}_{k=1}^K$, by drawing function sample paths $\tilde{f}^{(k)}$ from the GP posterior approximated by $q(f | \mathbf{x})$ and numerically maximizing each path to find $\mathbf{x}_*^{(k)}$.

Conditioning on $\mathbf{x}_*^{(k)}$ being the maximizer constrains the latent posterior at each candidate x through $f(x) \leq f(\mathbf{x}_*^{(k)})$, where $f(\mathbf{x}_*^{(k)})$ is the value the sample path attains at its optimum. The predictive term (10) uses the unconstrained posterior, so the acquisition measures the reduction in censored predictive entropy induced by the sampled optimum.

The expected conditional entropy is then approximated as:

$$\mathbb{E}_{p(\mathbf{x}|\mathcal{D})} H[y | \mathbf{x}, \mathbf{x}_*, \mathcal{D}] \approx \frac{1}{K} \sum_{k=1}^K H[y | \mathbf{x}, \mathbf{x}_*^{(k)}, \mathcal{D}]. \quad (12)$$

This formulation captures the information gain w.r.t. the global optimum by averaging the conditional entropy over the sampled locations of $\mathbf{x}_*$, while explicitly accounting for the boundaries and non-Gaussian nature of the censored likelihood in the predictive term.



(a) Metrics vs. censoring severity (mean and 95% c.i.). The upper row shows metrics w.r.t. censored observations, while the lower row shows them w.r.t. latent targets.

(b) Qualitative fits at censoring severity 0.3, seed=1.

Figure 3: **Model performance under varying data censoring.** *Include* uses all training points, passing censored observations to the learner. *Exclude* drops censored points. *Model* explicitly accounts for censoring in the likelihood; *Ignore* fits a standard Gaussian likelihood that treats clamped values as if uncensored. *Analytic* and *MC* refer to the estimator of $\mathbb{E}_q[\log p(y | f)]$.

6 EXPERIMENTS

We validate the following scenarios: the effect of censoring severity in Sec. 6.1, Bayesian Active Learning in Sec. 6.2, and Bayesian Optimization in Sec. 6.3. We also present real-world regression experiments in Appendix F. All experiments were implemented using GPyTorch [Gardner et al., 2018] and BOTorch [Balandat et al., 2020].

6.1 EFFECT OF CENSORING SEVERITY

We model the function introduced by Groot and Lucas [2012]: $f(x) = (g(x))^2 \sin(2g(x))$, where $g(x) = 6x - 2$. We censor data with increasing severity, using bounds that censor a specified fraction of the dataset.

Methods. We compare four ways to learn the true function:

1. **Model-Analytic (Include):** censored Variational Inference using our analytically derived ELBO.
2. **Model-MC (Include):** Montecarlo censored Variational Inference. We take 512 samples from the (Multivariate Normal) variational posterior to parameterize the Censored Gaussian likelihood.
3. **Ignore-Include (Gaussian):** naïvely treat clamped y as truth, using the default likelihood for GP regression.
4. **Ignore-Exclude (Truncated):** drop censored points.

Metrics. We compute the following on test-case data:

1. **Negative Log Predictive Density (NLPD):** we evaluate the posteriors under the *censored* predictive likelihood. This corresponds to passing the uncensored likelihood model through a clamping function.
2. **Mean Absolute Error (MAE):** this 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.

We compute the metrics in the following two cases:

1. **Observed-space metrics:** NLPD_c and MAE_c under the predictive likelihood against the clamped test targets. These metrics answer: how well do we model what we actually observe?
2. **Latent-space metrics:** NLPD_f and MAE_f against the true latent function values (only available in synthetic data). These metrics answer: how well do we recover the underlying signal despite censoring? These curves typically worsen with censoring severity; good censoring models degrade more slowly than baselines.

Results. Fig 3 shows the benefit of taking censoring into account. Ignore-Include (Gaussian) overfits to the empirical density at the censoring bounds, failing to recover the true structure beyond the bounds. It exhibits very low uncertainty around the censored regions due to the accumulated data

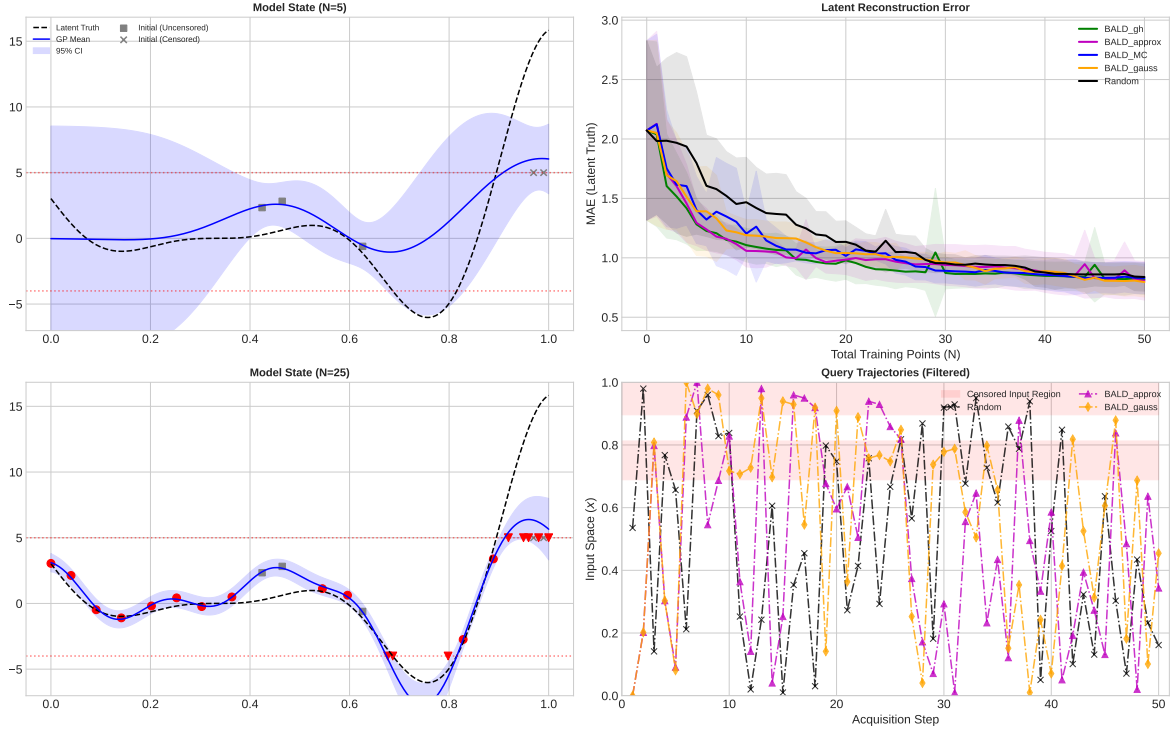


Figure 4: **Active Learning with Censored Observations.** (Rows 1-2) Evolution of the GP belief for the BALD_approx strategy. At $N = 5$, the model has high uncertainty (blue shading) in unobserved regions. By $N = 25$, the model has successfully identified the latent function (black dashed line), even in regions where observations are censored (red triangles indicate values clipped at boundaries $y = \pm 0.5$). (Row 3) Query trajectories show the sequential selection of input points $x \in [0, 1]$. Strategies like BALD_approx (purple triangles) actively sample near the decision boundaries of the censored regions (shaded red), whereas Random (black crosses) queries the space uniformly. (Row 4) Mean Absolute Error (MAE) relative to the latent truth. All AL strategies outperform random sampling, achieving lower error with fewer examples.

density. Our Model-Analytic (Include) likelihood shows a significant improvement in capturing the structure of the generative function. As expected with Montecarlo simulation, 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 with the analytic ELBO at significantly lower sample complexity.

In Appendix F.2, we extend the 1D severity study with a real-data severity sweep on Boston Housing dataset, comparing our VI against EP, LA and censoring-agnostic baselines across three censoring regimes ($\approx 3\%$, 26% , and 58%). The results confirm that the advantage of modelling censoring grows monotonically with severity, and that our analytical ELBO attains the best accuracy vs. compute trade-off among Gaussian approximations of the censored likelihood.

6.2 BAYESIAN ACTIVE LEARNING

We use GP regression with a censored likelihood to learn a latent function $f(x)$ with observations clipped to the interval $[-0.5, 0.5]$. The experiment is initialized with $N_{\text{init}} = 5$ random points, followed by a budget of $N_{\text{budget}} = 20$ acqui-

sition steps. At each step, the model selects the next query point x^* from a pool of 100 candidate locations in $[0, 1]$ by maximizing the acquisition score.

Methods. We compare four variants of the BALD objective against random sampling:

1. Analytical approximation (ours): BALD_approx
2. Gauss-Hermite quadrature: BALD_gh
3. Monte Carlo integration: BALD_MC
4. Gaussian approximation: BALD_gauss.

Recall, that C-BALD, [Hüttel et al., 2025], is a BALD score that uses MC evaluation. If we replace the BNN with GP and consider a special case where the censoring mechanism is known with probability 1, C-BALD reduces to our BALD_MC.

Results. The left column of Figure 4 illustrates the learning process for BALD_approx at the start ($N = 5$) and end ($N = 25$) of the budget. Initially, the model exhibits high uncertainty in unobserved regions. BALD_approx actively targets the boundaries of the censored regions, as shown the second row of the right column in Figure 4 by the red shaded bands. This allows the learner to reconstruct

Table 1: Simple regret (mean $\pm$ std. dev. over 10 independent seeds) on the Bayesian Optimization benchmark of Sec. 6.3, comparing Censored PES against a Censored GP with random acquisition and standard EI (Gaussian likelihood).

Strategy	Step 1	Step 5	Step 10	Step 15	Step 20	Step 25	Step 30
CensoredGP_PES	3.09 ± 2.30	2.40 ± 2.32	0.83 ± 1.52	0.58 ± 1.57	0.55 ± 1.58	0.09 ± 0.15	0.05 ± 0.08
CensoredGP_Random	3.04 ± 2.24	1.77 ± 1.85	0.99 ± 1.49	0.59 ± 0.93	0.15 ± 0.30	0.15 ± 0.30	0.15 ± 0.30
Standard_EI	2.73 ± 2.39	2.04 ± 2.57	1.21 ± 2.03	0.57 ± 1.57	0.50 ± 1.59	0.50 ± 1.59	0.05 ± 0.17

the latent function despite the information loss caused by clipping. The random strategy wastes queries on redundant regions that provide little information about the latent structure. The quantitative performance of different variants of BALD score is summarised in the first row of the right column in Figure 4. Despite the stochasticity induced by initializing different training pools across runs, all BALD variants demonstrate significantly faster convergence than random sampling.

Appendix F.3 extends this evaluation to Boston Housing dataset under the same censored active learning protocol. All BALD variants improve NLPD, RMSE, and MAE by 5–14% over random sampling at the final budget, and the closed-form BALD_approx of Lemma 4.3 is competitive with the more expensive MC estimator.

6.3 BAYESIAN OPTIMIZATION

We compare Censored Predictive Entropy Search (PES) against two baselines: a Censored GP with random acquisition, and standard Expected Improvement (EI) using a Gaussian likelihood that ignores censoring. We use GP regression with a censored likelihood to optimize a latent function $f(x)$ where observations were clipped to the interval $[C_L, C_U] = [-4.0, 5.0]$. We use the latent function, $f(x) = (6x - 2)^2 \sin(2(6x - 2))$, and corrupt observations with Gaussian noise ($\sigma_n = 0.8$) before censoring.

The experiment was initialized with $N_{\text{init}} = 5$ randomly selected points from a pool of 100 candidate locations in $[0, 1]$. We performed Bayesian Optimization for a budget of $N_{\text{budget}} = 30$ steps. At each step, the Censored_PES strategy selected the next query point x^* by maximizing the acquisition score defined in (10), which approximates the mutual information between the observation and the global optimizer using the derived censored entropy terms. The expected conditional entropy was approximated using $K = 10$ Thompson samples of the global optimizer x_* .

Metric. We use *Simple Regret*, defined as $R_t = f(x^*) - \max_{i=1\dots t} f(x_i)$, where $f(x^*)$ is the true global maximum of the latent function and $\max f(x_i)$ is the best latent value associated with any queried point up to step t . We averaged results over 10 independent seeds to account for the stochasticity of initialization and sampling.

Results. Table 1 shows that Censored PES converges faster

than both baselines, reaching a simple regret below 0.1 within 25 steps, versus 30 for Censored GP with random acquisition and a substantially slower, higher-variance decay for standard EI, which does not model the censoring mechanism. This confirms that explicitly accounting for censoring in the acquisition function accelerates convergence to the latent global maximum despite the partially observed objective.

7 CONCLUSION

We revisited Gaussian Process regression under value censoring and showed that key quantities for learning and decision making admit tractable forms. Concretely, we derived a numerically stable, closed-form ELBO for a Tobit (value-censored) likelihood, a closed-form predictive entropy, and an associated BALD objective that enables information-theoretic active learning and extends naturally to Bayesian optimization. Across synthetic stress tests and real tabular benchmarks with controlled censoring, the censored-likelihood GP improves observed-space metrics and typically matches or exceeds latent-space metrics relative to uncensored baselines, while avoiding the variance and wall-time overheads of Monte Carlo estimators.

Limitations. 1. Known, deterministic bounds: We assume censoring bounds (ℓ_i, u_i) are given and fixed at train/test time. Mis-specified bounds can potentially degrade calibration. A natural extension is to infer bounds (global or per-group) with hierarchical priors, or to handle partially known bounds via inequality constraints. **2. Homoscedastic noise:** We assume constant observation noise. Heteroscedastic or heavy-tailed noise can interact with censoring in ways a homoscedastic Gaussian cannot capture. **3. Myopic acquisitions:** Our BALD/PES formulations are one-step myopic and leverage local Gaussian structure. They come without global optimality guarantees. **4. Empirical validity:** Our real-data evaluations use synthetic censoring; although this isolates the effect of censoring, truly censored measurement pipelines may introduce additional biases (missing-not-at-random, protocol-specific noise). Validating on natively censored domains is an important next step.

References

- Milton Abramowitz and Irene Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1970.
- Stefan Aeberhard, Danny Coomans, and Olivier De Vel. Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition*, 27(8):1065–1077, 1994.
- Takeshi Amemiya. Tobit models: A survey. *Journal of econometrics*, 24(1-2):3–61, 1984.
- Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020. URL <http://arxiv.org/abs/1910.06403>.
- Marno Basson, Tobias Louw, and Theresa Smith. Variational tobit gaussian process regression. *Statistics and Computing*, 33(3), March 2023. ISSN 0960-3174. doi: 10.1007/s11222-023-10225-3.
- Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.
- Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. 2004.
- Emre Ertin. Gaussian process models for censored sensor readings. In *Statistical Signal Processing, IEEE/SP Workshop on*, pages 665–669, Los Alamitos, CA, USA, aug 2007. IEEE Computer Society. doi: 10.1109/SSP.2007.4301342.
- P. Friederichs and A. Hense. 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>.
- Daniele Gammelli, Inon Peled, Filipe Rodrigues, Dario Pacino, Hacı A. Kurtaran, and Francisco C. Pereira. Estimating latent demand of shared mobility through censored gaussian processes. *Transportation Research Part C: Emerging Technologies*, 120:102775, 2020. doi: <https://doi.org/10.1016/j.trc.2020.102775>. URL <https://www.sciencedirect.com/science/article/pii/S0968090X20306859>.
- Daniele Gammelli, Kasper Pryds Rolsted, Dario Pacino, and Filipe Rodrigues. Generalized multi-output gaussian process censored regression. *Pattern Recognition*, 129: 108751, 2022.
- Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, 2018.
- Perry Groot and Peter JF Lucas. Gaussian process regression with censored data using expectation propagation. 2012.
- David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.
- Marcelo Hartmann and Jarno Vanhatalo. Laplace approximation and natural gradient for gaussian process regression with heteroscedastic student-t model. *Statistics and Computing*, 29(4):753–773, 2019. doi: 10.1007/s11222-018-9836-0.
- James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, 2013.
- José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, volume 27, pages 918–926, 2014.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.
- Frederik Boe Hüttel, Christoffer Riis, Filipe Rodrigues, and Francisco Pereira. Bayesian active learning for censored regression. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2025, Porto, Portugal, September 15–19, 2025, Proceedings, Part II*, page 36–51. Springer-Verlag, 2025. ISBN 978-3-032-05980-2. doi: 10.1007/978-3-032-05981-9_3.
- Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. Bayesian optimization with censored response data. *arXiv*, 2013. doi: 10.48550/arxiv.1310.1947.
- Olav Kallenberg. *Foundations of Modern Probability*, volume 1. Springer, 3 edition, 2021.
- John P Klein, Melvin L Moeschberger, et al. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.
- Rui Li, S. T. John, and Arno Solin. Improving hyperparameter learning under approximate inference in gaussian process models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th*

- International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19632–19656. PMLR, 2023. URL <https://proceedings.mlr.press/v202/li23m.html>.
- Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
- Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.
- Tom Minka. Divergence measures and message passing. *Microsoft Research Technical Report*, MSR-TR-2005-173, 2005.
- Hoora Moradian, Denis Larocque, and François Bellavance. 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>.
- Chandra Nair, Balaji Prabhakar, and Devavrat Shah. On entropy for mixtures of discrete and continuous variables. *arXiv preprint cs/0607075*, 2006.
- Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078, 2008. URL <https://www.jmlr.org/papers/volume9/nickisch08a/nickisch08a.pdf>.
- Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009. URL http://www0.cs.ucl.ac.uk/staff/C.Archambeau/publ/neco_mo09_web.pdf.
- Liam Paninski. Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17(7):1480–1507, 2005.
- Tim Pearce, Jong-Hyeon Jeong, yichen jia, and Jun Zhu. Censored quantile regression neural networks for distribution-free survival analysis. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7450–7461. Curran Associates, Inc., 2022.
- Jean-Marc Robin. *Tobit model*, pages 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.
- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
- Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1015–1022, 2010.
- Emma Svensson, Hannah Rosa Friesacher, Susanne Winiwarter, Lewis Mervin, Adam Arany, and Ola Engkvist. Enhancing uncertainty quantification in drug discovery with censored regression labels. *Artificial Intelligence in the Life Sciences*, 2025. URL <https://www.sciencedirect.com/science/article/pii/S2667318525000042>.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933. doi: 10.1093/biomet/25.3-4.285. URL <https://www.jstor.org/stable/2332286>.
- Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.
- James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.
- Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with student-t likelihood. *Advances in Neural Information Processing Systems*, 22, 2009.
- Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

A MIXED CONTINUOUS AND DISCRETE MEASURES

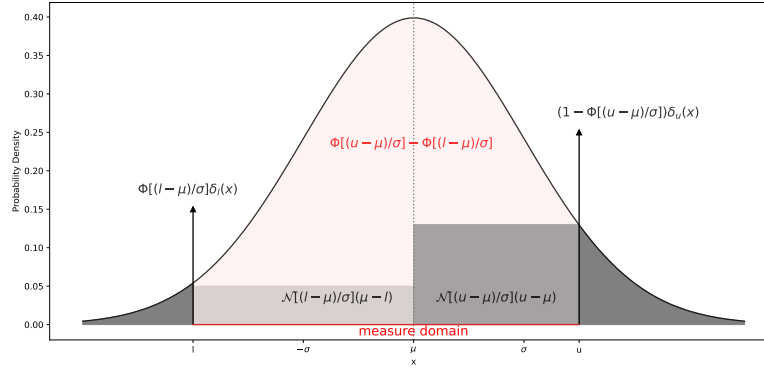


Figure 5: Censored Normal Distribution: the data observed from a normal distribution with values censored to lie in $[l, u]$. The Dirac delta terms capture the size of the censored tails of the underlying density. The depicted density corresponds to the data treatment where the events which happen outside the observation window (measure domain) are assigned the boundary value of this window. This does not change our perspective on how often the events occurs but it changes the odds of how often events occur at the boundaries.

Working with a mixed probability measure requires a different approach, as standard frameworks for purely discrete or purely continuous distributions are not directly applicable. We must treat the continuous regions, defined by a probability density, separately from the discrete point masses introduced by the censoring process. While other methods exist—for instance, [Nair et al., 2006] model this as a “mixed pair” to analyze the entropy rate of MCMC —our approach uses fundamental measure theory for a more rigorous treatment [Kallenberg, 2021].

A.1 CENSORED NORMAL DISTRIBUTION

The analysis of mixed probability measures requires a specialized framework. By the Lebesgue decomposition theorem, any such measure can be uniquely split into an absolutely continuous (or diffusive) part and a singular part. Here, the singular part consists of discrete point masses, or atoms.

To formalize this, we define a reference measure as $\rho = \lambda + \delta_l + \delta_u$, where λ denotes the Lebesgue measure and δ_z is Dirac measure located at atom z . The censored empirical measure ν_{emp}^c , associated with the real-world data, is absolutely continuous wrt ρ as anytime $\rho(A) = 0$, then also $\nu_{emp}^c(A) = 0$ for any $A \in \mathcal{A}$. The opposite does not hold and so the measures ν_{emp}^c and ν are not equivalent. e

Let the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ - algebra over the real line and P is the probability measure on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Recall the censored measure restricted on the compact interval with the normal density $\phi(x; \mu, \sigma)$ and mixed-in two probability masses p_l and p_u at its boundaries. Its Radon-Nykodym derivative wrt to the Lebesgue measure with two Dirac’s atoms at l and u is given by:

$$p(x; \mu, \sigma, l, u) = \Phi\left(\frac{l-\mu}{\sigma}\right)\delta_l(x) + \phi(x; \mu, \sigma)\mathbb{I}_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)]\delta_u(x), \quad (13)$$

where $\mathbb{I}_{(l,u)}$ is the identifier 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 normal distribution is not absolutely continuous wrt to the censored normal distribution as it assigns non-zero mass to the tails, which the censored normal measures as a null set.

Moments of Censored Normal Distribution

In the following we provide the formulas for the moments of the censored normal distribution, the discussion on the properties which these formulas imply and their derivation.

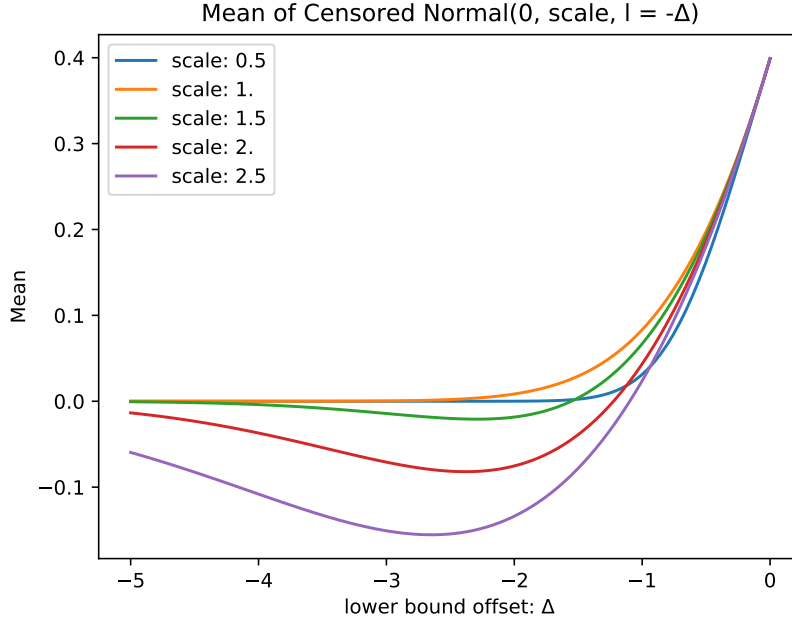


Figure 6: Mean of the asymmetric one-sided censored normal distribution: the impact of scale of the underlying normal distribution centred at 0 on the mean of the lower-bounded censored normal distribution $p(x; 0, \sigma, -\Delta)$ given by the relation: $\sigma\phi(-\frac{\Delta}{\sigma}) - \Delta\Phi(-\frac{\Delta}{\sigma})$.

The mean of the censored normal distribution $p(x; \mu, \sigma, l, u)$ is given:

$$\begin{aligned} \mathbb{E}p(x; \mu, \sigma, l, u) = & \mu[\Phi(\frac{u-\mu}{\sigma}) - \Phi(\frac{l-\mu}{\sigma})] - \sigma[\phi(\frac{u-\mu}{\sigma}) - \phi(\frac{l-\mu}{\sigma})] + \\ & + l\Phi(\frac{l-\mu}{\sigma}) + u[1 - \Phi(\frac{u-\mu}{\sigma})], \end{aligned} \quad (14)$$

and the variance of the censored normal distribution $p(x; \mu, \sigma, l, u)$ is given:

$$\begin{aligned} \text{Var } p(x; \mu, \sigma, l, u) = & (\mu^2 + \sigma^2) \left[\Phi(\frac{u-\mu}{\sigma}) - \Phi(\frac{l-\mu}{\sigma}) \right] - 2\mu\sigma \left[\phi(\frac{u-\mu}{\sigma}) - \phi(\frac{l-\mu}{\sigma}) \right] - \\ & - \sigma^2 \left[\frac{u-\mu}{\sigma} \phi(\frac{u-\mu}{\sigma}) - \frac{l-\mu}{\sigma} \phi(\frac{l-\mu}{\sigma}) \right] + l^2 \Phi(\frac{l-\mu}{\sigma}) + u^2 [1 - \Phi(\frac{u-\mu}{\sigma})] - \mu_c^2. \end{aligned} \quad (15)$$

where we denote $\mu_c = \mathbb{E}p(x; \mu, \sigma, l, u)$, $\phi(\mu, \sigma)$ and $\Phi(\mu, \sigma)$ the density and the cumulative distributional function (CDF) of the normal distribution

We see that both moments of the original distribution are scaled by the probability mass of the observation window and compensated for asymmetry plus the moments of the atoms.

In case of the mean of the censored normal, the compensation captures the difference of the density values of the censoring boundaries scaled by the standard deviation, see the second term in (14). The two terms on the second line of (14) are the mean values of the Bernoulli distribution taking the boundary values with probabilities of the collapsed tails.

In case of variance, we compensate with the difference of rectangles between the censoring boundary and centre of the mass of uncensored distribution with the height corresponding to the density.

In the symmetric case with censoring boundaries $l = \mu - \Delta$ and $u = \mu + \Delta$, the mean value equals to the mean of the uncensored normal distribution μ :

$$\begin{aligned} \mathbb{E}p(x; \mu, \sigma, \mu - \Delta, \mu + \Delta) = & \\ = & \mu[\Phi(\frac{\Delta}{\sigma}) - \Phi(-\frac{\Delta}{\sigma})] - \sigma[\phi(\frac{\Delta}{\sigma}) - \phi(-\frac{\Delta}{\sigma})] + (\mu - \Delta)\Phi(-\frac{\Delta}{\sigma}) + (\mu + \Delta)[1 - \Phi(\frac{\Delta}{\sigma})] = \\ = & \mu[\Phi(\frac{\Delta}{\sigma}) - \Phi(-\frac{\Delta}{\sigma}) + \Phi(-\frac{\Delta}{\sigma}) + 1 - \Phi(\frac{\Delta}{\sigma})] - \Delta[\Phi(-\frac{\Delta}{\sigma}) - \Phi(\frac{\Delta}{\sigma})] \\ = & \mu. \end{aligned}$$

For the variance of the censored normal distribution we have:

$$\begin{aligned}
\text{Var } p(\mu, \sigma, \mu - \Delta, \mu + \Delta) &= \\
&= (\mu^2 + \sigma^2) [\Phi(\frac{\Delta}{\sigma}) - \Phi(-\frac{\Delta}{\sigma})] - 2\mu\sigma [\phi(\frac{\Delta}{\sigma}) - \phi(-\frac{\Delta}{\sigma})] - \sigma\Delta [\phi(\frac{\Delta}{\sigma}) + \phi(-\frac{\Delta}{\sigma})] + \\
&+ \Delta^2\Phi(-\frac{\Delta}{\sigma}) + \Delta^2[1 - \Phi(\frac{\Delta}{\sigma})] - \mu^2 \\
&= (\mu^2 + \sigma^2) [\Phi(\frac{\Delta}{\sigma}) - \Phi(-\frac{\Delta}{\sigma})] - 2\Delta\sigma\phi(\frac{\Delta}{\sigma}) + 2\Delta^2\Phi(-\frac{\Delta}{\sigma}) - \mu^2,
\end{aligned}$$

which demonstrates the impact of the symmetric censoring on the variance of the original distribution. In particular, if we censor far ends of the distribution symmetrically, the impact of the censoring on the variance σ^2 of the underlying normal distribution is minimal.

On the other hand for one-sided censoring, e.g. with lower bound censoring $l = \mu - \Delta$ and $u \rightarrow \infty$, the mean depends only on the size of censoring and the variance of the underlying distribution:

$$\begin{aligned}
\mathbb{E}p(x; \mu, \sigma, \mu - \Delta) &= -\mu\Phi(-\frac{\Delta}{\sigma}) + \sigma\phi(-\frac{\Delta}{\sigma}) + (\mu - \Delta)\Phi(-\frac{\Delta}{\sigma}) \\
&= \sigma\phi(-\frac{\Delta}{\sigma}) - \Delta\Phi(-\frac{\Delta}{\sigma}).
\end{aligned} \tag{16}$$

For special case of $\mu = 0$, $l = 0$ and $u \rightarrow \infty$, we recover one half of the mean of half-normal distribution: $\sigma/\sqrt{2\pi}$. Note that the half-normal distribution is a truncated normal distribution and so it needs to be re-scaled. The half-censored normal distribution places atom of the size 0.5 to point 0 and keeps the remaining half of the original normal density unchanged.

Derivation of the Mean

To arrive at the formula for the mean of the censored normal distribution we need to compute the following integral:

$$\begin{aligned}
\mathbb{E}p(x; \mu, \sigma, l, u) &= \int_{\mathbb{R}} x \left[\Phi(\frac{l-\mu}{\sigma})\delta_l(x) + \phi(x; \mu, \sigma)\mathbb{I}_{(l,u)}(x) + [1 - \Phi(\frac{u-\mu}{\sigma})]\delta_u(x) \right] dx = \\
&= l\Phi(\frac{l-\mu}{\sigma}) + \int_l^u x\phi(x; \mu, \sigma)dx + u[1 - \Phi(\frac{u-\mu}{\sigma})].
\end{aligned} \tag{17}$$

From here we see that to even numerically integrate this density, we need to integrate the area within the observation window (l, u) and add measure values at the boundary points at l and u separately.

Let us evaluate the middle term in the relation (17). Recall $\left[-e^{-\frac{z^2}{2}}\right] dz = ze^{-\frac{z^2}{2}}$ and substitute $z = \frac{x-\mu}{\sigma}$:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi\sigma^2}} \int_l^u xe^{-\frac{(x-\mu)^2}{2\sigma^2}} dx &= \frac{\mu}{\sqrt{2\pi\sigma^2}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} e^{-\frac{z^2}{2}} \sigma dz + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} \sigma ze^{-\frac{z^2}{2}} \sigma dz = \\
&= \mu[\Phi(\frac{u-\mu}{\sigma}) - \Phi(\frac{l-\mu}{\sigma})] - \sigma[\phi(\frac{u-\mu}{\sigma}) - \phi(\frac{l-\mu}{\sigma})].
\end{aligned} \tag{18}$$

Putting it all together we arrive at the formula:

$$\begin{aligned}
\mathbb{E}p(x; \mu, \sigma, l, u) &= \mu[\Phi(\frac{u-\mu}{\sigma}) - \Phi(\frac{l-\mu}{\sigma})] - \sigma[\phi(\frac{u-\mu}{\sigma}) - \phi(\frac{l-\mu}{\sigma})] + \\
&+ l\Phi(\frac{l-\mu}{\sigma}) + u[1 - \Phi(\frac{u-\mu}{\sigma})].
\end{aligned} \tag{19}$$

Derivation of the Variance

Let us denote the mean of the censored normal distribution $p(x; \mu, \sigma, l, u)$ as: $\mu_c = \mathbb{E}p(x; \mu, \sigma, l, u)$. To arrive at the formula for the variance of the censored normal distribution we need to compute the following integral:

$$\begin{aligned}
 \text{Var } p(x; \mu, \sigma, l, u) &= \int_{\mathbb{R}} (x - \mu_c)^2 \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \mathbb{I}_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] dx = \\
 &= \int_{\mathbb{R}} (x^2 - 2x\mu_c + \mu_c^2) \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \mathbb{I}_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] dx = \\
 &= \int_{\mathbb{R}} x^2 \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \mathbb{I}_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] dx - \mu_c^2 = \\
 &= l^2 \Phi\left(\frac{l-\mu}{\sigma}\right) + \int_l^u x^2 \phi(x; \mu, \sigma) dx + u^2 [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] - \mu_c^2.
 \end{aligned} \tag{20}$$

For computing the second term we use substitution $z = \frac{x-\mu}{\sigma}$. Because $\left[-ze^{-\frac{z^2}{2}}\right] dz = z^2 e^{-\frac{z^2}{2}} - e^{-\frac{z^2}{2}}$ and $\left[-e^{-\frac{z^2}{2}}\right] dz = ze^{-\frac{z^2}{2}}$ we get:

$$\begin{aligned}
 \frac{1}{\sqrt{2\pi}\sigma^2} \int_l^u x^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx &= \frac{1}{\sqrt{2\pi}\sigma^2} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} (\sigma^2 z^2 + 2\mu\sigma z + \mu^2) e^{-\frac{z^2}{2}} \sigma dz = \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} z^2 e^{-\frac{z^2}{2}} dz + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} ze^{-\frac{z^2}{2}} dz + \frac{\mu^2}{\sqrt{2\pi}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz = \\
 &= \frac{\mu^2 + \sigma^2}{\sqrt{2\pi}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz - \frac{2\mu\sigma}{\sqrt{2\pi}} \left[e^{-\frac{z^2}{2}} \right]_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} - \frac{\sigma^2}{\sqrt{2\pi}} \left[ze^{-\frac{z^2}{2}} \right]_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} = \\
 &= (\mu^2 + \sigma^2) \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] - 2\mu\sigma \left[\phi\left(\frac{u-\mu}{\sigma}\right) - \phi\left(\frac{l-\mu}{\sigma}\right) \right] - \sigma^2 \left[\frac{u-\mu}{\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) - \frac{l-\mu}{\sigma} \phi\left(\frac{l-\mu}{\sigma}\right) \right].
 \end{aligned} \tag{21}$$

Putting it all together we get:

$$\begin{aligned}
 \text{Var } p(x; \mu, \sigma, l, u) &= (\mu^2 + \sigma^2) \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] - 2\mu\sigma \left[\phi\left(\frac{u-\mu}{\sigma}\right) - \phi\left(\frac{l-\mu}{\sigma}\right) \right] - \\
 &\quad - \sigma^2 \left[\frac{u-\mu}{\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) - \frac{l-\mu}{\sigma} \phi\left(\frac{l-\mu}{\sigma}\right) \right] + l^2 \Phi\left(\frac{l-\mu}{\sigma}\right) + u^2 [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] - \mu_c^2.
 \end{aligned} \tag{22}$$

B INFORMATION THEORY

Consider two measures ν and ρ defined on some measurable space $(\Omega, \mathcal{A})$ with ν being absolutely continuous wrt to a reference measure ρ . The entropy of ν wrt to reference measure ρ is defined as:

$$H\left(\frac{d\nu}{d\rho}\right) = \int_{\Omega} \left[\log \frac{d\nu}{d\rho} \right] \frac{d\nu}{d\rho} \rho(\omega), \tag{23}$$

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

The relative entropy, also referred to as 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:

$$D\left(\frac{d\nu_1}{d\rho} \parallel \frac{d\nu_2}{d\rho}\right) = \int_{\Omega} \left[\log \frac{d\nu_2}{d\nu_1} \right] \frac{d\nu_1}{d\rho} \rho(\omega). \tag{24}$$

Similarly we can define the cross-entropy or the probability measures ν_1, ν_2 :

$$H\left(\frac{d\nu_1}{d\rho} \parallel \frac{d\nu_2}{d\rho}\right) = \int_{\Omega} \left[\log \frac{d\nu_2}{d\rho} \right] \frac{d\nu_1}{d\rho} \rho(\omega). \tag{25}$$

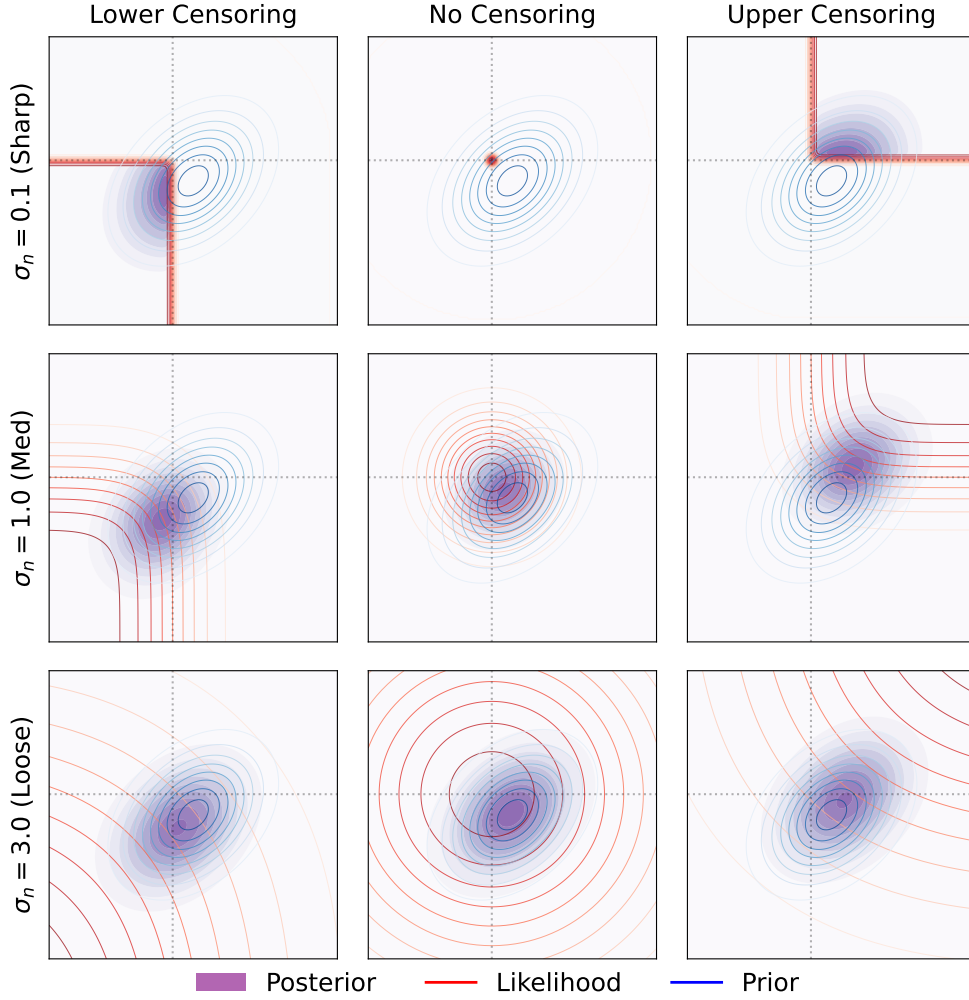


Figure 7: **Impact of Censoring and Noise:** we show the GP posterior distribution over two latent function values, $f_1 = f(\mathbf{x}_1)$ and $f_2 = f(\mathbf{x}_2)$, corresponding to inputs $\mathbf{x}_1 = (1, 1)'$ and $\mathbf{x}_2 = (-1, -1)'$. The observations are $y_1 = -0.5$ and $y_2 = 0.5$. The rows illustrate different censoring regimes: lower censoring ($c = -1$) in the top row ($f_i \leq y_i$), standard Gaussian regression ($c = 0$) in the middle ($f_i \approx y_i$), and upper censoring ($c = 1$) in the bottom row ($f_i \geq y_i$). Columns correspond to increasing likelihood noise scales $\sigma_n \in \{0.1, 1.0, 3.0\}$. Blue contours represent the GP prior with RBF kernel with lengthscale $l = 2.0$, while red contours indicate the likelihood constraints.

B.1 INFORMATION THEORY FOR MIXTURES OF DISCRETE AND CONTINUOUS VARIABLES

The concept of entropy for mixtures of discrete and continuous distributions has been studied e.g. in [Nair et al., 2006] in connection with entropy rate of continuous-time Markov chains.

The main theoretical justification comes from the Radon-Nikodym decomposition of the measures which characterises the decomposition of the measure into its diffusive part and the discrete part. So it allows to mix in the atoms which corresponds to the singular measures wrt Lebesgue measure without jeopardizing the existence of the integrals. This does not pose any issue for the existence of the information-theoretical quantities as long as the absolute continuity of measure ν wrt ρ is satisfied.

B.2 INFORMATION THEORETICAL QUANTITIES FOR CENSORED NORMAL DISTRIBUTION

Next we present and discuss the analytical formula for entropy of the normal distribution with the location parameter μ and the scale parameter σ censored from the bottom with the parameter l and top with the parameter u and $l < u$. It is given by:

$$H[p(\mu, \sigma, l, u)] = \log[\sqrt{2\pi e}\sigma] \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] - \frac{1}{2} \left[\frac{u-\mu}{\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) - \frac{l-\mu}{\sigma} \phi\left(\frac{l-\mu}{\sigma}\right) \right] - \Phi\left(\frac{l-\mu}{\sigma}\right) \log \Phi\left(\frac{l-\mu}{\sigma}\right) - \Phi\left(\frac{\mu-u}{\sigma}\right) \log \Phi\left(\frac{\mu-u}{\sigma}\right). \quad (26)$$

Recalling Figure 5, the first term in the formula corresponds to the entropy of the normal distribution scaled down by the actual probability mass of the observation window. The second term in the formula captures the asymmetry of the distribution which happens due to censoring. The last two terms are the entropy of the probability masses of the collapsed tails. For censoring boundaries $l = \mu - \Delta$ and $u = \mu + \Delta$, we can simplify the entropy formula as:

$$H[p(x; \mu, \sigma, \mu - \Delta, \mu + \Delta)] = \log[\sqrt{2\pi e}\sigma] \left[\Phi\left(\frac{\Delta}{\sigma}\right) - \Phi\left(-\frac{\Delta}{\sigma}\right) \right] - \frac{\Delta}{\sigma} \phi\left(\frac{\Delta}{\sigma}\right) - \Phi\left(-\frac{\Delta}{\sigma}\right) \log \Phi\left(-\frac{\Delta}{\sigma}\right) - \Phi\left(\frac{\Delta}{\sigma}\right) \log \Phi\left(\frac{\Delta}{\sigma}\right).$$

This formula also gives us insight that if we censor symmetrically far ends of the distribution, the impact of the censoring on the entropy of the distribution is very small.

The entropy for censoring applied only from the bottom with parameter l is given by:

$$H[p(\mu, \sigma, l)] = \log[\sqrt{2\pi e}\sigma] \left[1 - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] + \frac{l-\mu}{2\sigma} \phi\left(\frac{l-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \log \Phi\left(\frac{l-\mu}{\sigma}\right). \quad (27)$$

and the entropy for censoring applied only from the top with parameter u is:

$$H[p(\mu, \sigma, u)] = \log[\sqrt{2\pi e}\sigma] \Phi\left(\frac{u-\mu}{\sigma}\right) - \frac{u-\mu}{2\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{u-\mu}{\sigma}\right) \log \Phi\left(\frac{u-\mu}{\sigma}\right). \quad (28)$$

We are also interested in deriving the KL divergence of the normal distribution with the density $\phi(m, s)$ and censored normal distribution $p(\mu, \sigma, l, u)$ with the underlying normal density $\phi(\mu, \sigma)$. 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 . 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 while the opposite does not hold.

The KL divergence of the normal distribution with the density $\phi(m, s)$ and censored normal distribution $p(\mu, \sigma, l, u)$ with the underlying normal density $\phi(\mu, \sigma)$ is:

$$KL(\phi(m, s) || p(\mu, \sigma, l, u)) = -\log[\sqrt{2\pi e}s] + \left\{ \frac{1}{2} \log[2\pi\sigma^2] - \frac{(\mu-m)^2 + s^2}{2\sigma^2} \right\} [\Phi\left(\frac{u-m}{s}\right) - \Phi\left(\frac{l-m}{s}\right)] - \left\{ \log \Phi\left(\frac{\mu-u}{\sigma}\right) + \frac{(u+m-2\mu)s}{2\sigma^2} \right\} \phi\left(\frac{u-m}{s}\right) - \left\{ \log \Phi\left(\frac{l-\mu}{\sigma}\right) - \frac{(l+m-2\mu)s}{2\sigma^2} \right\} \phi\left(\frac{l-m}{s}\right).$$

Derivation of the Entropy

Let us compute the entropy of the censored normal distribution:

$$\begin{aligned} -H[p(x; \mu, \sigma, l, u)] &= \int_{\mathbb{R}} \log p(x; \mu, \sigma, l, u) \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \mathbb{I}_{(l, u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] dx \\ &= \left[\log p(x; \mu, \sigma, l, u) \Phi\left(\frac{l-\mu}{\sigma}\right) \right]_{x=l} + \int_l^u \log p(x; \mu, \sigma, l, u) \phi(x; \mu, \sigma) dx + \left[\log p(x; \mu, \sigma, l, u) [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \right]_{x=u} \\ &= \Phi\left(\frac{l-\mu}{\sigma}\right) \log \Phi\left(\frac{l-\mu}{\sigma}\right) + \int_l^u \phi(x; \mu, \sigma) \log \phi(x; \mu, \sigma) dx + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \log [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)]. \end{aligned} \quad (29)$$

We can further simplify the above expression. Let us compute the integral term in the middle:

$$I \equiv \int_l^u \phi(x; \mu, \sigma) \log \phi(x; \mu, \sigma) dx = -\frac{1}{2} \log[2\pi\sigma^2] \int_l^u \phi(x; \mu, \sigma) dx - \frac{1}{\sqrt{2\pi\sigma^2}} \int_l^u \frac{(x-\mu)^2}{2\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx. \quad (30)$$

The first term is the difference of the normal CDF at the transformed boundaries. For computing the second term we use substitution $z = \frac{x-\mu}{\sigma}$:

$$I = -\frac{1}{2} \log[2\pi\sigma^2] \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] - \frac{\sigma}{2\sqrt{2\pi\sigma^2}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} z^2 e^{-\frac{z^2}{2}} dz \quad (31)$$

Because $\left[-ze^{-\frac{z^2}{2}} \right] dz = z^2 e^{-\frac{z^2}{2}} - e^{-\frac{z^2}{2}}$, we can compute the second term of the integral I in the closed form:

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} z^2 e^{-\frac{z^2}{2}} dz &= \frac{1}{\sqrt{2\pi}} \int_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz - \frac{1}{\sqrt{2\pi}} \left[ze^{-\frac{z^2}{2}} \right]_{\frac{l-\mu}{\sigma}}^{\frac{u-\mu}{\sigma}} \\ &= \Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) - \frac{u-\mu}{\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) + \frac{l-\mu}{\sigma} \phi\left(\frac{l-\mu}{\sigma}\right). \end{aligned} \quad (32)$$

Putting it all together, we arrive at the closed-form analytical expression of the censored normal entropy:

$$\begin{aligned} H[p(x; l, u)] &= -\Phi\left(\frac{l-\mu}{\sigma}\right) \log \Phi\left(\frac{l-\mu}{\sigma}\right) - \Phi\left(\frac{\mu-u}{\sigma}\right) \log \Phi\left(\frac{\mu-u}{\sigma}\right) + \frac{1}{2} \log[2\pi\sigma^2] \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] \\ &\quad + \frac{1}{2} \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] - \frac{1}{2} \left[\frac{u-\mu}{\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) - \frac{l-\mu}{\sigma} \phi\left(\frac{l-\mu}{\sigma}\right) \right] \\ &= \log[\sqrt{2\pi e}\sigma] \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] - \frac{1}{2} \left[\frac{u-\mu}{\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) - \frac{l-\mu}{\sigma} \phi\left(\frac{l-\mu}{\sigma}\right) \right] \\ &\quad - \Phi\left(\frac{l-\mu}{\sigma}\right) \log \Phi\left(\frac{l-\mu}{\sigma}\right) - \Phi\left(\frac{\mu-u}{\sigma}\right) \log \Phi\left(\frac{\mu-u}{\sigma}\right). \end{aligned} \quad (33)$$

We can now derive the entropy for the one-sided censoring. We also show that in the special case we retain the entropy of the underlying normal distribution. To compute the entropy of the one-sided upper censoring we relax the lower censoring bound:

$$\lim_{l \rightarrow -\infty} H[p(x; l, u)] = \log[\sqrt{2\pi e}\sigma] \Phi\left(\frac{u-\mu}{\sigma}\right) - \frac{u-\mu}{2\sigma} \phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{\mu-u}{\sigma}\right) \log \Phi\left(\frac{\mu-u}{\sigma}\right), \quad (34)$$

because $\lim_{x \rightarrow -\infty} \Phi(x) \log \Phi(x) = 0$. The order of the integral of the exponential of the square decays to 0 much faster than logarithm or any power law decays to $-\infty$. Analogically, we can derive the entropy of the lower bounded censoring bound of the normal distribution:

$$\lim_{u \rightarrow \infty} H[p(x; l, u)] = \log[\sqrt{2\pi e}\sigma] \left[1 - \Phi\left(\frac{l-\mu}{\sigma}\right) \right] + \frac{l-\mu}{2\sigma} \phi\left(\frac{l-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \log \Phi\left(\frac{l-\mu}{\sigma}\right). \quad (35)$$

To obtain the entropy of the underlying normal distribution we relax the remaining boundary again:

$$\lim_{u \rightarrow \infty} \lim_{l \rightarrow -\infty} H[p(x; l, u)] = \lim_{l \rightarrow -\infty} \lim_{u \rightarrow \infty} H[p(x; l, u)] = \log[\sqrt{2\pi e}\sigma]. \quad (36)$$

It is easy to verify that the equity holds also for the pairs: $(u \rightarrow -\infty, l \rightarrow \infty)$ and $(l \rightarrow \infty, u \rightarrow -\infty)$.

Derivation of the Cross Entropy and KL Divergence

Recall that KL-divergence between the two probability distributions P and Q can be computed as a difference between the cross-entropy of Q wrt to P and the entropy of P .

Let us derive the (negative) cross entropy of the censored normal distribution $p(\mu, \sigma, l, u)$ wrt to the normal distribution with

a density $\phi(m, s)$:

$$\begin{aligned}
& \mathbb{E}_{\phi(m,s)} \log p(\mu, \sigma, l, u) \\
&= \int_{\mathbb{R}} \phi(x; m, s) \log \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \delta_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] dx \\
&= \left[\phi(x; m, s) \log \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \delta_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] \right]_{x=l} \\
&\quad + \int_l^u \phi(x; m, s) \log \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \delta_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] dx \\
&\quad + \left[\phi(x; m, s) \log \left[\Phi\left(\frac{l-\mu}{\sigma}\right) \delta_l(x) + \phi(x; \mu, \sigma) \delta_{(l,u)}(x) + [1 - \Phi\left(\frac{u-\mu}{\sigma}\right)] \delta_u(x) \right] \right]_{x=u} \\
&= \phi\left(\frac{l-m}{s}\right) \log[\Phi\left(\frac{l-\mu}{\sigma}\right)] + \int_l^u \phi(x; m, s) \log \phi(x; \mu, \sigma) dx + \phi\left(\frac{u-m}{s}\right) \log[\Phi\left(\frac{\mu-u}{\sigma}\right)].
\end{aligned} \tag{37}$$

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

$$I \equiv \int_l^u \phi(x; m, s) \log \phi(x; \mu, \sigma) dx = -\frac{1}{2} \log[2\pi\sigma^2] \int_l^u \phi(x; m, s) dx - \frac{1}{\sqrt{2\pi}s^2} \int_l^u \frac{(x-m)^2}{2s^2} e^{-\frac{(x-m)^2}{2s^2}} dx. \tag{38}$$

The first term is the difference between the two normal CDF values. To evaluate the second integral we start with using the following trick: $[(x-m) + (m-\mu)]^2 = (x-m)^2 + 2x(m-\mu) + \mu^2 - m^2$ and the result from (32):

$$\begin{aligned}
& \frac{1}{\sqrt{2\pi}s^2} \int_l^u \frac{(x-m)^2}{2s^2} e^{-\frac{(x-m)^2}{2s^2}} dx = \frac{s^2}{\sigma^2 \sqrt{2\pi}s^2} \int_l^u \frac{(x-m)^2 + 2x(m-\mu) + \mu^2 - m^2}{2s^2} e^{-\frac{(x-m)^2}{2s^2}} dx \\
&= \frac{s^2}{2\sigma^2} [\Phi\left(\frac{u-m}{s}\right) - \Phi\left(\frac{l-m}{s}\right) - \frac{u-m}{s} \phi\left(\frac{u-m}{s}\right) + \frac{l-m}{s} \phi\left(\frac{l-m}{s}\right)] + \frac{2(m-\mu)}{2\sigma^2 \sqrt{2\pi}s^2} \int_l^u x e^{-\frac{(x-m)^2}{2s^2}} dx + \\
&\quad + \frac{\mu^2 - m^2}{2\sigma^2} [\Phi\left(\frac{u-m}{s}\right) - \Phi\left(\frac{l-m}{s}\right)].
\end{aligned} \tag{39}$$

The last integral, which we need to evaluate, corresponds to the first moment of the censored normal distribution with the location parameter m and the scale parameter s excluding the atoms the boundaries. Recall $\left[-e^{-\frac{z^2}{2}}\right] dz = ze^{-\frac{z^2}{2}}$ and substitute $z = \frac{x-m}{s}$

$$\begin{aligned}
& \frac{1}{\sqrt{2\pi}s^2} \int_l^u x e^{-\frac{(x-m)^2}{2s^2}} dx = \frac{s}{\sqrt{2\pi}} \int_{\frac{l-m}{s}}^{\frac{u-m}{s}} z e^{-\frac{z^2}{2}} dz + \frac{m}{\sqrt{2\pi}s^2} \int_{\frac{l-m}{s}}^{\frac{u-m}{s}} e^{-\frac{z^2}{2}} dz \\
&= s[\phi\left(\frac{l-m}{s}\right) - \phi\left(\frac{u-m}{s}\right)] + m[\Phi\left(\frac{u-m}{s}\right) - \Phi\left(\frac{l-m}{s}\right)].
\end{aligned} \tag{40}$$

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

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

The formula for the cross entropy is:

$$\begin{aligned}
& -\mathbb{E}_{\phi(m,s)} \log p(\mu, \sigma, l, u) \\
&= -\phi\left(\frac{l-m}{s}\right) \log[\Phi\left(\frac{l-\mu}{\sigma}\right)] - I - \phi\left(\frac{u-m}{s}\right) \log[\Phi\left(\frac{\mu-u}{\sigma}\right)] \\
&= \left\{ \frac{1}{2} \log[2\pi\sigma^2] + \frac{(\mu-m)^2 + s^2}{2\sigma^2} \right\} [\Phi\left(\frac{u-m}{s}\right) - \Phi\left(\frac{l-m}{s}\right)] \\
&\quad - \left\{ \log \Phi\left(\frac{\mu-u}{\sigma}\right) + \frac{(u+m-2\mu)s}{2\sigma^2} \right\} \phi\left(\frac{u-m}{s}\right) \\
&\quad - \left\{ \log \Phi\left(\frac{l-\mu}{\sigma}\right) - \frac{(l+m-2\mu)s}{2\sigma^2} \right\} \phi\left(\frac{l-m}{s}\right)
\end{aligned} \tag{41}$$

To arrive at KL-divergence we subtract the entropy of the Gaussian $\phi(m, s)$ from the cross-entropy above:

$$\begin{aligned} KL(\phi(m, s)||p(l, u)) &= -\log[\sqrt{2\pi}es] + \left\{\frac{1}{2}\log[2\pi\sigma^2] + \frac{(\mu-m)^2+s^2}{2\sigma^2}\right\}[\Phi(\frac{u-m}{s}) - \Phi(\frac{l-m}{s})] \\ &\quad - \left\{\log\Phi(\frac{\mu-u}{\sigma}) + \frac{(u+m-2\mu)s}{2\sigma^2}\right\}\phi(\frac{u-m}{s}) \\ &\quad - \left\{\log\Phi(\frac{l-\mu}{\sigma}) - \frac{(l+m-2\mu)s}{2\sigma^2}\right\}\phi(\frac{l-m}{s}) \end{aligned} \quad (42)$$

C BALD SCORE

For an input x , let the latent approximate posterior be

$$q(f|x, \mathcal{D}) = \mathcal{N}(f_x|m_x, s_x^2)$$

Observations arise from a Gaussian with homoscedastic noise $\sigma_y > 0$. For simplicity we assume that the observations are censored to an interval $[l, u]$ (allowing $l = -\infty$ or $u = +\infty$). The Mutual Information can be computed as predictive entropy less the expected conditional entropy with respect to the posterior:

$$\begin{aligned} I(y; f|x, \mathcal{D}) &= H[y|x, \mathcal{D}] - \mathbb{E}_{p(f|x, \mathcal{D})}[H(y|f, x)] \approx H[y|x, \mathcal{D}] - \mathbb{E}_{\theta \sim q(f|x, \mathcal{D})}[H(y|f, \theta, x)] \\ &\approx H\left(\int \mathcal{CN}(f_x, \sigma_y^2; l, u) \mathcal{N}(f_x|m_x, s_x^2) df_x\right) - \int H(\mathcal{CN}(f_x, \sigma_y^2; l, u)) \mathcal{N}(f_x|m_x, s_x^2) df_x. \end{aligned} \quad (43)$$

PREDICTIVE ENTROPY

First, we analyze the predictive distribution. The convolution of the Gaussian posterior with the censored Gaussian likelihood results in a Censored Normal distribution with inflated variance $s_x^2 + \sigma_y^2$. The predictive entropy is derived by expanding the integral over the mixed measure, i.e. continuous density plus discrete masses. Let us use the formula (26) on the mean of the predictive posterior:

$$\begin{aligned} H &\left(\int \Phi(\frac{l-f_x}{\sigma_y}) \mathcal{N}(f_x|m_x, s_x^2) df_x + \int \mathcal{N}(f_x|0, \sigma_y^2) \mathcal{N}(f_x|m_x, s_x^2) df_x + \int [1 - \Phi(\frac{f_x-u}{\sigma_y})] \mathcal{N}(f_x|m_x, s_x^2) df_x\right) \\ &= H\left(\Phi\left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}}\right) + \mathcal{N}(m_x, \sigma_y^2 + s_x^2) + 1 - \Phi\left(\frac{m_x-u}{\sqrt{\sigma_y^2+s_x^2}}\right)\right) = H(\mathcal{CN}(m_x, \sqrt{s_x^2 + \sigma_y^2}; l, u)) \\ &= \log[\sqrt{2\pi}e(s_x^2 + \sigma_y^2)] \left[\Phi\left(\frac{u-m_x}{\sqrt{\sigma_y^2+s_x^2}}\right) - \Phi\left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}}\right)\right] - \frac{1}{2} \left[\frac{u-m_x}{\sqrt{\sigma_y^2+s_x^2}} \phi\left(\frac{u-m_x}{\sqrt{\sigma_y^2+s_x^2}}\right) - \frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}} \phi\left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}}\right)\right] \\ &\quad - \Phi\left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}}\right) \log \Phi\left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}}\right) - \Phi\left(\frac{m_x-u}{\sqrt{\sigma_y^2+s_x^2}}\right) \log \Phi\left(\frac{m_x-u}{\sqrt{\sigma_y^2+s_x^2}}\right). \end{aligned} \quad (44)$$

EXPECTED CONDITIONAL ENTROPY

Next, we focus on the second term. The expected conditional entropy requires integrating the entropy of the likelihood over the posterior distribution of f :

$$\begin{aligned} \int H(\mathcal{CN}(f_x, \sigma_y^2; l, u)) \mathcal{N}(f_x|m_x, s_x^2) df_x &= \log \sqrt{2\pi}e\sigma_y \left[\int \Phi(\frac{u-f_x}{\sigma_y}) \mathcal{N}(f_x|m_x, s_x^2) df_x - \int \Phi(\frac{l-f_x}{\sigma_y}) \mathcal{N}(f_x|m_x, s_x^2) df_x\right] \\ &\quad - \frac{1}{2} \left[\int \frac{u-f_x}{\sigma_y} \mathcal{N}(u|f_x, \sigma_y^2) \mathcal{N}(f_x|m_x, s_x^2) df_x - \int \frac{l-f_x}{\sigma_y} \mathcal{N}(l|f_x, \sigma_y^2) \mathcal{N}(f_x|m_x, s_x^2) df_x\right] \\ &\quad - \int \Phi(\frac{l-f_x}{\sigma_y}) \log \Phi(\frac{l-f_x}{\sigma_y}) \mathcal{N}(f_x|m_x, s_x^2) df_x - \int \Phi(\frac{f_x-u}{\sigma_y}) \log \Phi(\frac{f_x-u}{\sigma_y}) \mathcal{N}(f_x|m_x, s_x^2) df_x. \end{aligned} \quad (45)$$

Derivation of the Conditional Asymmetry Expectation

We derive the analytical form for the expected asymmetry term involving the upper bound u :

$$I = \int \frac{u-f_x}{\sigma_y} \mathcal{N}(u|f_x, \sigma_y^2) \mathcal{N}(f_x|m_x, s_x^2) df_x. \quad (46)$$

First, we apply the **product of Gaussians** identity to the integrand. The product of the likelihood and the posterior is:

$$\mathcal{N}(u | f_x, \sigma_y^2) \mathcal{N}(f_x | m_x, s_x^2) = \mathcal{N}(u | m_x, \sigma_y^2 + s_x^2) \cdot \mathcal{N}(f_x | \mu_c, \Sigma_c), \quad (47)$$

where $V^2 = \sigma_y^2 + s_x^2$ is the total predictive variance, and the conditional parameters are:

$$\mu_c = \frac{us_x^2 + m_x\sigma_y^2}{s_x^2 + \sigma_y^2}, \quad \Sigma_c = \left(\frac{1}{\sigma_y^2} + \frac{1}{s_x^2} \right)^{-1}. \quad (48)$$

Substituting this back into the integral, we can factor out the predictive density $\mathcal{N}(u | m_x, V^2)$ which does not depend on f_x :

$$I = \frac{1}{\sigma_y} \mathcal{N}(u | m_x, V^2) \int (u - f_x) \mathcal{N}(f_x | \mu_c, \Sigma_c) df_x = \frac{1}{\sigma_y} \mathcal{N}(u | m_x, V^2) \left[u - \underbrace{\mathbb{E}_{\mathcal{N}(\mu_c, \Sigma_c)}[f_x]}_{\mu_c} \right]. \quad (49)$$

The term inside the bracket simplifies algebraically:

$$u - \mu_c = u - \frac{us_x^2 + m_x\sigma_y^2}{s_x^2 + \sigma_y^2} = \frac{u(s_x^2 + \sigma_y^2) - (us_x^2 + m_x\sigma_y^2)}{s_x^2 + \sigma_y^2} = \frac{\sigma_y^2(u - m_x)}{V^2}. \quad (50)$$

Now, we substitute this back into the expression for I . We also use the identity $\mathcal{N}(u | m_x, V^2) = \frac{1}{V} \phi\left(\frac{u - m_x}{V}\right)$:

$$I = \frac{1}{\sigma_y} \left[\frac{1}{V} \phi\left(\frac{u - m_x}{V}\right) \right] \left[\frac{\sigma_y^2(u - m_x)}{V^2} \right] = \frac{\sigma_y}{V^2} \left[\frac{u - m_x}{V} \phi\left(\frac{u - m_x}{V}\right) \right]. \quad (51)$$

This derivation shows that the expected conditional asymmetry term is scaled by a factor of $\frac{\sigma_y^2}{V^2}$. In the final BALD score, this term is subtracted from the predictive asymmetry (which effectively has a scaling of 1), yielding the final coefficient:

$$1 - \frac{\sigma_y^2}{\sigma_y^2 + s_x^2} = \frac{s_x^2}{\sigma_y^2 + s_x^2}. \quad (52)$$

So we get for the **expected conditional entropy**:

$$\begin{aligned} \mathbb{E}[H(y|f)] &= \int H(\mathcal{CN}(f_x, \sigma_y; l, u)) \mathcal{N}(f_x | m_x, s_x^2) df_x \\ &= \log(\sqrt{2\pi}e\sigma_y) \left[\Phi\left(\frac{u - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) - \Phi\left(\frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) \right] \\ &\quad - \frac{1}{2} \frac{\sigma_y^2}{\sigma_y^2 + s_x^2} \left[\frac{u - m_x}{\sqrt{\sigma_y^2 + s_x^2}} \phi\left(\frac{u - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) - \frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}} \phi\left(\frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) \right] \\ &\quad - \mathbb{E}_{\mathcal{N}(z|0,1)} \left[\Phi\left(\frac{l - m_x + s_x z}{\sigma_y}\right) \log \Phi\left(\frac{l - m_x + s_x z}{\sigma_y}\right) + \Phi\left(\frac{m_x - u + s_x z}{\sigma_y}\right) \log \Phi\left(\frac{m_x - u + s_x z}{\sigma_y}\right) \right]. \end{aligned} \quad (53)$$

Putting it all together, we get the corrected BALD score:

$$\begin{aligned} BALD(x) &= \frac{1}{2} \log\left(1 + \frac{s_x^2}{\sigma_y^2}\right) \left[\Phi\left(\frac{u - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) - \Phi\left(\frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) \right] \\ &\quad - \frac{1}{2} \frac{s_x^2}{\sigma_y^2 + s_x^2} \left[\frac{u - m_x}{\sqrt{\sigma_y^2 + s_x^2}} \phi\left(\frac{u - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) - \frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}} \phi\left(\frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) \right] \\ &\quad + \mathbb{E}_{\mathcal{N}(z|0,1)} \left[\Phi\left(\frac{l - m_x + s_x z}{\sigma_y}\right) \log \Phi\left(\frac{l - m_x + s_x z}{\sigma_y}\right) + \Phi\left(\frac{m_x - u + s_x z}{\sigma_y}\right) \log \Phi\left(\frac{m_x - u + s_x z}{\sigma_y}\right) \right] \\ &\quad - \Phi\left(\frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) \log \Phi\left(\frac{l - m_x}{\sqrt{\sigma_y^2 + s_x^2}}\right) - \Phi\left(\frac{m_x - u}{\sqrt{\sigma_y^2 + s_x^2}}\right) \log \Phi\left(\frac{m_x - u}{\sqrt{\sigma_y^2 + s_x^2}}\right). \end{aligned} \quad (54)$$

The remaining expectation terms can be approximated efficiently using Gaussian convolution [Houlsby et al., 2011] as derived in the following subsection, or evaluated using e.g. trapezoid rule or Gauss-Hermite polynomials.

UNCENSORED SANITY CHECK

With $l = -\infty$, $u = +\infty$, the above formula reduces to

$$H[Y|x, \mathcal{D}] = \frac{1}{2} \log(2\pi e(s_x^2 + \sigma_y^2)), \quad \mathbb{E}[H(Y|f, x)] = \frac{1}{2} \log(2\pi e\sigma_y^2)$$

hence

$$BALD(x) = \frac{1}{2} \log \left(1 + \frac{s_x^2}{\sigma_y^2} \right),$$

the standard Gaussian Mutual Information.

C.1 DERIVATION OF THE APPROXIMATION OF THE INTEGRAL TERMS

In the calculation of the expected conditional entropy for the Censored BALD score, we are left with the intractable integrals of the form:

$$I = \mathbb{E}_{z \sim \mathcal{N}(m_x, s_x^2)} [\Phi(z) \log \Phi(z)], \quad (55)$$

where $z = (l - f)/\sigma_y$ (or $(f - u)/\sigma_y$) is a standardized latent variable. The function $g(z) = \Phi(z) \log \Phi(z)$ describes the entropy contribution of the atoms.

Following [Houlsby et al., 2011], who proposed to use Taylor expansion on the Bernoulli entropy with probit link function, we leverage the fact that $g(z)$ is smooth, negative, unimodal, and decays to zero as $|z| \rightarrow \infty$. We approximate it using a scaled, unnormalized Gaussian (squared exponential) function:

$$\Phi(z) \log \Phi(z) \approx \hat{g}(z) = -A \exp \left(-\frac{(z - \mu_0)^2}{2\sigma_0^2} \right). \quad (56)$$

This substitution transforms the intractable integral into a Gaussian convolution, which admits an analytical solution.

Determination of Constants

The constants are chosen to match the geometric properties of the original function $g(z)$ and to minimize integration error.

Amplitude (A) and Location (μ_0): We find the stationary point by setting the derivative to zero:

$$\frac{d}{dz} (\Phi(z) \log \Phi(z)) = \phi(z) (1 + \log \Phi(z)) = 0. \quad (57)$$

Since $\phi(z) > 0$, the condition implies $\log \Phi(z) = -1$, or $\Phi(z) = 1/e$.

- The location is $\mu_0 = \Phi^{-1}(1/e) \approx -0.337$.
- The peak amplitude is $g(\mu_0) = (1/e) \log(1/e) = -1/e$. Thus, $A = 1/e \approx 0.3679$.

Width (σ_0): While the function fit yields $\sigma \approx 0.90$, optimizing specifically to minimize the squared error of the *integral* across a range of variances yields a slightly adjusted width:

$$\sigma_0 \approx 0.9794 \quad (58)$$

Derivation of the Closed-Form Integral

Substituting the ansatz $\hat{g}(z)$ into the expectation integral I :

$$I \approx \int_{-\infty}^{\infty} \left[-A \exp \left(-\frac{(z - \mu_0)^2}{2\sigma_0^2} \right) \right] \frac{1}{\sqrt{2\pi}b} \exp \left(-\frac{(z - a)^2}{2b^2} \right) dz, \quad (59)$$

where $a = (l - m_x)/\sigma_y$ and $b = s_x/\sigma_y$.

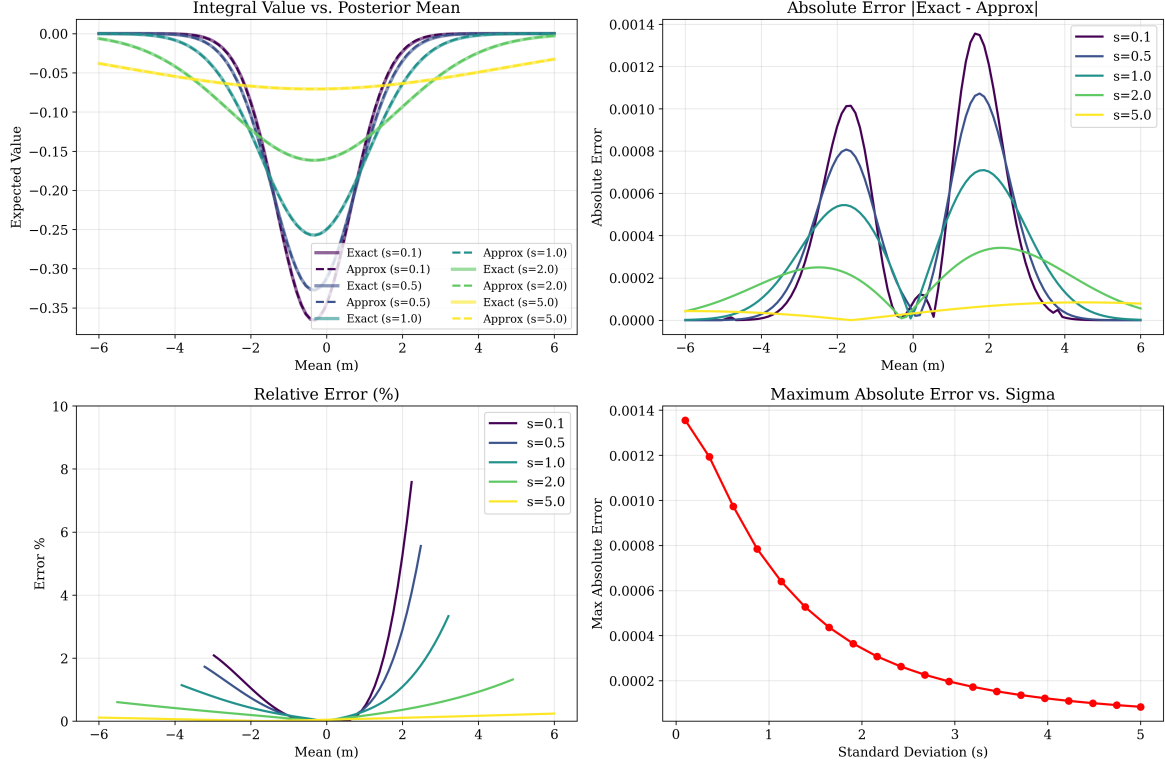


Figure 8: **Accuracy of the Analytical Approximation of the Integral Terms.** (Top Left) Comparison of the closed-form Gaussian convolution (dashed) against exact numerical integration (solid) for the term $\mathbb{E}_{\mathcal{N}(z|\mu, \sigma^2)}[\Phi(z) \log \Phi(z)]$. (Top Right) Absolute approximation error across the latent domain. (Bottom Left) Relative error percentage. (Bottom Right) Maximum absolute error as a function of the predictive standard deviation σ . The approximation becomes increasingly accurate as model uncertainty grows.

We apply the **product of Gaussians** identity. The product of two exponential quadratic forms is proportional to a single exponential:

$$\exp\left(-\frac{(z - \mu_0)^2}{2\sigma_0^2}\right) \exp\left(-\frac{(z - a)^2}{2b^2}\right) = S \cdot \exp\left(-\frac{(z - \mu_{\text{new}})^2}{2\sigma_{\text{new}}^2}\right), \quad (60)$$

where the scaling factor S (the value of the convolution at the means) is:

$$S = \exp\left(-\frac{(a - \mu_0)^2}{2(b^2 + \sigma_0^2)}\right). \quad (61)$$

Recognizing that the integral of the remaining unnormalized Gaussian term is $\sqrt{2\pi\sigma_{\text{new}}^2}$ (or alternatively using the standard expectation identity for Gaussian kernels), we arrive at the final analytical approximation:

$$\mathbb{E}_{z \sim \mathcal{N}(m_x, s_x^2)}[\Phi(z) \log \Phi(z)] \approx -\frac{A\sigma_0}{\sqrt{b^2 + \sigma_0^2}} \exp\left(-\frac{(b - \mu_0)^2}{2(b^2 + \sigma_0^2)}\right) = -\frac{A\sigma_y\sigma_0}{\sqrt{s_x^2 + \sigma_y^2\sigma_0^2}} \exp\left(-\frac{(l - m_x - \sigma_y\mu_0)^2}{2(s_x^2 + \sigma_y^2\sigma_0^2)}\right) \quad (62)$$

where the fitted constants are $A \approx 0.3679$, $\mu_0 \approx -0.337$, and $\sigma_0 \approx 0.9794$.

C.2 APPROXIMATION ACCURACY

To justify the use of the analytical Gaussian approximation for the expected conditional entropy, we evaluated its accuracy against exact numerical integration. The core difficulty lies in computing the expectation $\mathbb{E}_{z \sim \mathcal{N}(m, \sigma^2)}[\Phi(z) \log \Phi(z)]$. We

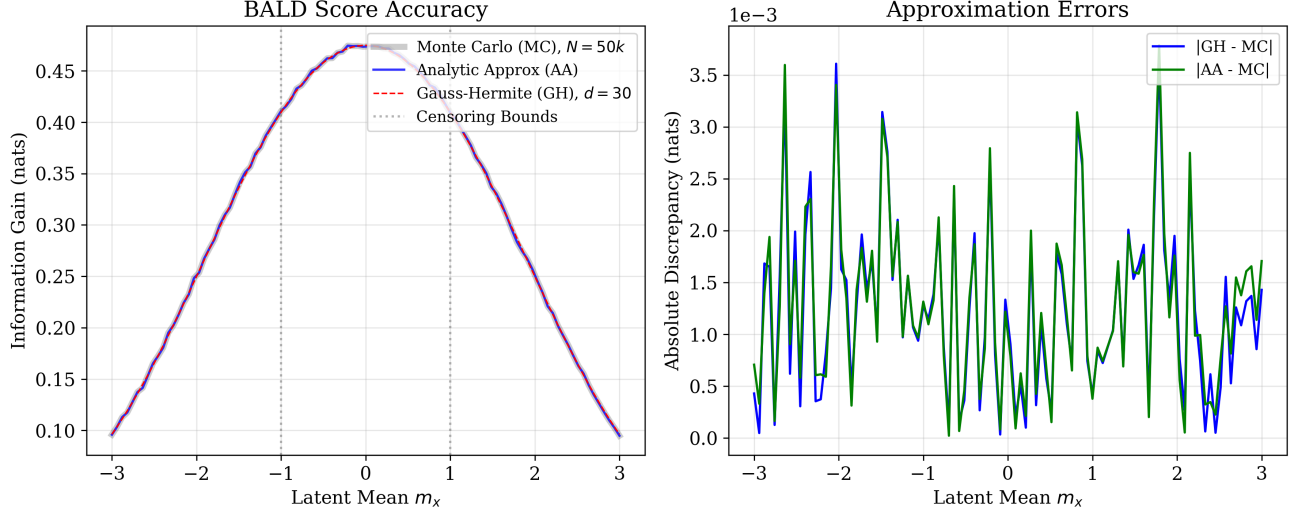


Figure 9: **Verification of the Analytical BALD Approximation.** Comparison between the proposed Gauss-Hermite quadrature approximation (degree $d = 30$) and a high-fidelity Monte Carlo baseline ($N = 50,000$ samples) for the Censored BALD score. The left panel shows the score profiles for a fixed variance $s_x^2 = 1.0$ and censoring bounds $[-1, 1]$. The right panel displays the absolute residual error, demonstrating that the approximation error is negligible ($< 10^{-3}$ nats) across the domain.

approximated the integrand $\Phi(z) \log \Phi(z)$ as a scaled Gaussian function, allowing the integral to be solved in closed form as a convolution.

We compared this closed-form solution against ground-truth numerical quadrature (SciPy `quad`) across a wide range of posterior means $m \in [-6, 6]$ and standard deviations $\sigma \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$. As shown in Figure 8, the approximation (dashed lines) closely tracks the exact value (solid lines) across all regimes. The absolute error is highest in regions of low uncertainty ($\sigma = 0.1$) where the non-Gaussian skew of the integrand is most pronounced, yet it remains sufficiently low (< 0.02) for active learning purposes. As uncertainty increases ($\sigma \geq 1.0$), the convolution smooths out these discrepancies, rendering the approximation nearly indistinguishable from the exact integral.

To validate the precision of the proposed Gauss-Hermite quadrature approximation for the BALD score, we compared it against a high-fidelity Monte Carlo baseline. The experiment evaluated the score across a sweep of latent posterior means $m_x \in [-3, 3]$, representing regions ranging from the uncensored center to the heavily censored tails. We fixed the model uncertainty at $s_x^2 = 1.0$ and observation noise at $\sigma_y^2 = 0.5$, with symmetric censoring bounds $[l, u] = [-1, 1]$.

The BALD score is decomposed into the exact predictive entropy and the expected conditional entropy. For the latter term, the Monte Carlo baseline estimated the integral using $N = 50,000$ samples drawn from the latent posterior $q(f|x)$. The proposed method utilized Gauss-Hermite quadrature with degree $d = 30$. As illustrated in Figure 9, the quadrature method achieves near-perfect alignment with the Monte Carlo baseline and same holds for the approximation formula. The maximum absolute discrepancy was observed to be less than 10^{-3} nats, confirming that deterministic quadrature provides a computationally efficient yet highly accurate alternative to stochastic sampling for the censored information gain.

C.3 BALD APPROXIMATE FORMULA

Substituting the derived Gaussian approximations from previous part, we get:

$$\begin{aligned}
BALD(x) = & \frac{1}{2} \log \left(1 + \frac{s_x^2}{\sigma_y^2} \right) \left[\Phi \left(\frac{u-m_x}{\sqrt{\sigma_y^2+s_x^2}} \right) - \Phi \left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}} \right) \right] \\
& - \frac{1}{2} \frac{s_x^2}{\sigma_y^2+s_x^2} \left[\frac{u-m_x}{\sqrt{\sigma_y^2+s_x^2}} \phi \left(\frac{u-m_x}{\sqrt{\sigma_y^2+s_x^2}} \right) - \frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}} \phi \left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}} \right) \right] \\
& - \Phi \left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}} \right) \log \Phi \left(\frac{l-m_x}{\sqrt{\sigma_y^2+s_x^2}} \right) - \Phi \left(\frac{m_x-u}{\sqrt{\sigma_y^2+s_x^2}} \right) \log \Phi \left(\frac{m_x-u}{\sqrt{\sigma_y^2+s_x^2}} \right) \\
& - \frac{A\sigma_y}{\sqrt{s_x^2+\sigma_y^2}} \left[\exp \left\{ -\frac{(l-m_x-\sigma_y^2\mu_0)^2}{2(s_x^2+\sigma_y^2\sigma_0^2)} \right\} - \exp \left\{ -\frac{(u-m_x-\sigma_y^2\mu_0)^2}{2(s_x^2+\sigma_y^2\sigma_0^2)} \right\} \right],
\end{aligned} \tag{63}$$

where $A = 1/e \approx 0.368$, $\mu_0 = \Phi^{-1}(1/e) \approx -0.337$ and fitted $\sigma_0 \approx 0.9794$.

D INFORMATION-THEORETIC DESIGN OF EXPERIMENT WITH CENSORED LIKELIHOOD

D.1 PROOF OF LEMMA 4.1

We need to verify the assumptions of Theorem 1 in Paninski [2005] for the case of Gaussian Processes with censored likelihoods.

Proof. We aim to show that the posterior mass concentrates on the true function f_0 as $N \rightarrow \infty$. The proof relies on the Doob's Martingale Convergence Theorem and the properties of the information gain objective.

Let $p_0(f)$ be the prior density. Consider the posterior ratio of the true parameter f_0 to any alternative $f \in \Theta$:

$$R_N(f) = \frac{p_N(f)}{p_N(f_0)} = \frac{p_0(f)}{p_0(f_0)} \prod_{i=1}^N \frac{p(y_i | x_i, f)}{p(y_i | x_i, f_0)}. \tag{64}$$

By Condition (C3), $p_0(f_0) > 0$, ensuring this ratio is well-defined. Taking the logarithm and analyzing the expected behavior under the true distribution $p(y | x, f_0)$, we observe that for any input sequence $\{x_i\}$, the log-likelihood ratio process behaves as a random walk with negative drift equal to the Kullback-Leibler divergence:

$$\mathbb{E}_{y \sim f_0} \left[\log \frac{p(y | x, f)}{p(y | x, f_0)} \right] = -D_{\text{KL}}(f_0; f | x). \tag{65}$$

Standard Bayesian consistency results (Doob's theorem) guarantee that if we sample x such that the accumulated divergence grows to infinity, the posterior mass on any $f \neq f_0$ vanishes ($R_N(f) \rightarrow 0$) almost surely.

The critical active learning step is to ensure that the chosen inputs x_i actually accumulate this divergence. The algorithm selects $x_{N+1} = \arg \max_{x \in \mathcal{X}} I(f; y | x, \mathcal{D}_N)$. Condition (C4) states that for any incorrect $f \neq f_0$, there exists an x^* such that $D_{\text{KL}}(f_0; f | x^*) > 0$. Condition (C2) (Lipschitz continuity of the censored likelihood) ensures that the information gain $I(x)$ is a continuous function of x . Combined with Condition (C1) of the domain $\mathcal{X}$, this continuity implies that if the posterior has not converged (i.e., there is significant mass on $f \neq f_0$), the maximum information gain is strictly bounded away from zero:

$$\max_{x \in \mathcal{X}} I(f; y | x, \mathcal{D}_N) \geq \delta > 0. \tag{66}$$

This ensures the algorithm cannot get “stuck” querying uninformative points.

Assume, for the sake of contradiction, that the posterior does not converge to f_0 . This implies there exists a set of incorrect functions $S_\epsilon = \{f : \rho(f, f_0) > \epsilon\}$ that retains non-negligible posterior mass as $N \rightarrow \infty$. However, as established in Step 2, the greedy strategy continuously selects x to maximize the divergence between the current posterior and the truth. By the chain rule of mutual information, the cumulative information gain is the reduction in posterior entropy. Since the total entropy is bounded (due to C1), the algorithm must eventually exhaust the potential information gain. The only state where the expected information gain is zero is when the posterior $p_N(f)$ is concentrated entirely on functions indistinguishable from f_0 . Given (C4), this implies concentration on f_0 itself.

Thus, the posterior probability mass of any neighborhood of f_0 converges to 1 in probability. $\square$

D.2 PROOF OF LEMMA 4.2

The proof demonstrates the asymptotic normality of the active learning posterior for the Gaussian Process projection, followed by the consistency of the MAP estimator and the efficiency of the covariance matrix. We rely on the asymptotic theory of information-theoretic experimental design as detailed in Paninski [2005], specifically adapting Theorem 1 and Corollary 1 to our Censored GP framework.

Proof. Standard Gaussian Processes define a prior over an infinite-dimensional function space. However, by evaluating the GP at a finite set of N query points $\mathbf{x}_1, \dots, \mathbf{x}_N$, we project the model onto a finite-dimensional Euclidean space $\mathbb{R}^N$. The prior distribution over this finite projection, $\mathbf{f}_N = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)]^\top$, is a multivariate Normal distribution: $p_0(\mathbf{f}_N) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{NN})$. Because $\mathbf{K}_{NN}$ is strictly positive definite for distinct inputs (assuming a standard kernel like RBF or Matérn), the prior is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^N$, and its density is bounded and strictly positive on any compact subset of $\mathbb{R}^N$.

Lemma 4.1 establishes that the posterior $p(\mathbf{f}_N | \mathcal{D}_N)$ is consistent, meaning it concentrates its mass in an infinitesimally small neighborhood around the true values $\mathbf{f}_{0,N}$. We now analyze the local shape of the posterior in this neighborhood. By assumption, the censoring link function Φ is twice continuously differentiable (C^2). Consequently, the log-likelihood function $\ell(\mathbf{f}_N) = \sum_{i=1}^N \log p(y_i | \mathbf{f}_N(\mathbf{x}_i))$ is also C^2 . Because the estimate is consistent, we can perform a Taylor expansion of the log-posterior around the truth $\mathbf{f}_{0,N}$. The curvature of the log-likelihood is governed by the empirical Fisher Information matrix, $\mathcal{I}_N(\mathbf{f}_N)$. The C^2 smoothness ensures that $\mathcal{I}_N$ is continuous and well-defined. As $N \rightarrow \infty$, the likelihood term dominates the bounded prior. By the adaptive Bernstein-von Mises theorem [Paninski, 2005, Theorem 1], the posterior distribution converges in total variation to a multivariate Gaussian whose covariance is the inverse of the Fisher Information matrix:

$$\lim_{N \rightarrow \infty} \int |p(\mathbf{f}_N | \mathcal{D}_N) - \mathcal{N}(\mathbf{f}_N | \mathbf{f}_{0,N}, \Sigma_N)| d\mathbf{f}_N = 0, \quad (67)$$

where $\Sigma_N = \mathcal{I}_N^{-1}$.

The Maximum A Posteriori (MAP) estimator is defined as the mode of the posterior distribution: $\hat{\mathbf{f}}_{\text{MAP}} = \arg \max_{\mathbf{f}_N} p(\mathbf{f}_N | \mathcal{D}_N)$. Because the prior density $p_0(\mathbf{f}_N)$ is bounded and continuous, and because the posterior converges in total variation to a Normal distribution $\mathcal{N}(\mathbf{f}_{0,N}, \Sigma_N)$ (which is unimodal and symmetric), the mode of the posterior asymptotically aligns with the mean of the Gaussian limit. Therefore, $\hat{\mathbf{f}}_{\text{MAP}} \rightarrow \mathbf{f}_{0,N}$ almost surely, and its asymptotic sampling distribution is exactly $\mathcal{N}(\mathbf{f}_{0,N}, \Sigma_N)$.

Finally, we evaluate the efficiency of the covariance matrix Σ_N . The information-maximization strategy (BALD) selects query points $\mathbf{x}_{N+1}$ that greedily maximize the mutual information $I(\mathbf{f}_N; y | x)$. In the asymptotic Gaussian regime, the differential entropy of the posterior $\mathcal{N}(\mathbf{f}_{0,N}, \Sigma_N)$ is analytically defined as:

$$H(p(\mathbf{f}_N | \mathcal{D}_N)) \approx \frac{1}{2} \log |(2\pi e)^N \Sigma_N|. \quad (68)$$

Because mutual information is the expected reduction in entropy ($I = H_{\text{prior}} - \mathbb{E}[H_{\text{posterior}}]$), an active learning strategy that maximizes mutual information is asymptotically equivalent to sequentially minimizing the log-determinant of the posterior covariance matrix, $\log |\Sigma_N|$. Since the objective is globally minimized by the information-theoretic strategy relative to drawing inputs from a fixed, non-adaptive sampling distribution, it strictly follows that the resulting volume of the confidence ellipsoid is upper-bounded by that of random sampling:

$$|\Sigma_{\text{info}}| \leq |\Sigma_{\text{iid}}|. \quad (69)$$

This inequality demonstrates that the information-theoretic design achieves the theoretically tightest confidence bounds (D-optimality) for the projected function values. $\square$

E EXPECTED LOG LIKELIHOOD COMPUTATION

Suppose we have a probabilistic model (e.g. a Gaussian process, or a Bayesian linear model) which marginally yields a latent Gaussian distribution over f , which we will denote $q(f|m, s) = \mathcal{N}(f|m, s^2)$. To construct an ELBO we typically will need to compute

$$\begin{aligned} \mathbb{E}_{q(f|m, s)}[\log p(y|f, \sigma)] &= \delta_{(\ell, u)}(y) \mathbb{E}_f[\log \tilde{p}(y|f, \sigma)] \\ &\quad + \delta_\ell(y) \mathbb{E}_f[\log p_\ell(f, \sigma)] + \delta_u(y) \mathbb{E}_f[\log p_u(f, \sigma)]. \end{aligned} \quad (70)$$

This expression above is for a single datapoint y . We will typically compute the sum of this across many y_i values, each with a particular over f_i ; for each y_i , only a single term will contribute to the overall ELBO.

The middle term, for observations where $y \in (\ell, u)$, is a straightforward result:

$$\begin{aligned}\mathbb{E}_{q(f|m,s)}[\log \tilde{p}(y|f, \sigma)] \\ &= \mathbb{E}_{\mathcal{N}(f|m,s^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-f)^2 \right] \\ &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2ym + m^2 + s^2).\end{aligned}$$

The two atoms at the censored boundaries are trickier.

E.1 NUMERICAL INSTABILITY OF LOG CDF

Before considering specific approximations or estimators to the expected log CDF in the above equation, we first should consider how to evaluate the integrand itself. Unfortunately, simply evaluating $\log \Phi(z)$ using standard numerical packages, with $\Phi(z)$ as an intermediate result, leads to underflow for even moderately negative values. Accordingly, we would like to directly evaluate the log CDF in a numerically stable way, or evaluate a lower bound of it if that is not feasible. Most existing published bounds for the CDF of a normal are expressed in terms of the Gaussian Q-function [Abramowitz and Stegun, 1970]: $Q(z) = 1 - \Phi(z) = \Phi(-z)$.

A well-known lower bound is $Q(z) > \frac{z}{1+z^2} \phi(z)$, over the range $z \in (0, \infty)$ which with some manipulation gives a lower bound on $\log \Phi(z)$ as

$$\begin{aligned}\log \Phi(z) &> \log(-z) - \log(1+z^2) + \log \phi(-z) \\ &= \log(-z) - \log(1+z^2) - \frac{z^2}{2} - \frac{1}{2} \log 2\pi.\end{aligned}$$

This is theoretically valid over the left half of the real line, $z \in (-\infty, 0)$, and only loosely near zero. However, at single-precision the “standard” calculation of $\log \Phi(z)$ only begins to underflow (with the derivative becoming numerically unstable) around $z < -5$, at which point the lower bound is fairly tight. Consequently, we choose to use the lower bound instead of the naïve $\log \Phi(z)$ computation for any sufficiently small values of z .

E.2 NUMERIC QUADRATURE FOR ATOM PROBABILITIES

The atom probabilities are univariate integrals under a Gaussian, so this is quite appropriate for simple numeric quadrature. We first express the atom probabilities in terms of integrals with respect to a standard Gaussian, with

$$\begin{aligned}\mathbb{E}_{q(f|m,s)}[\log p_\ell(f, \sigma)] &= \mathbb{E}_{\mathcal{N}(f|m,s^2)} \left[\log \Phi\left(\frac{\ell-f}{\sigma}\right) \right] \\ &= \mathbb{E}_{\mathcal{N}(z|0,1)} \left[\log \Phi\left(\frac{\ell-m}{\sigma} + \frac{s}{\sigma}z\right) \right] \\ \mathbb{E}_{q(f|m,s)}[\log p_u(f, \sigma)] &= \mathbb{E}_{\mathcal{N}(z|0,1)} \left[\log \Phi\left(\frac{m-u}{\sigma} + \frac{s}{\sigma}z\right) \right].\end{aligned}$$

Since a standard normal pdf is less than 1×10^{-8} for values $\|z\| > 6$, we can restrict the numeric integration to a fixed region $z \in [-6, 6]$. So, each of the two atom probabilities $f(z) = \log \Phi(a + bz)$ can be expressed by the integral:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} f(z) dz \approx \frac{1}{\sqrt{2\pi}} \int_{-6}^6 e^{-\frac{z^2}{2}} f(z) dz$$

We show that the simple trapezoidal rule works sufficiently and allows for a batch computation. See the Supplementary for detailed discussion and tests of the implementation and its robustness.

Table 2: Performance metrics on real-world regression benchmarks (Mean $\pm$ Std. Err. across 5 bootstrap iterations).

Dataset	Method	NLPD_f $\downarrow$	MAE_f $\downarrow$	NLPD_c $\downarrow$	MAE_c $\downarrow$
Boston Housing ($N_D=11$)	Model-Analytic (Ours)	2.80 ± 0.02	2.53 ± 0.06	1.00 ± 0.07	1.37 ± 0.11
	Ignore-Include (Gaussian)	3.56 ± 0.08	2.91 ± 0.03	1.63 ± 0.10	1.87 ± 0.06
Credit Risk ($N_D=8$)	Model-Analytic (Ours)	3.61 ± 0.03	7.20 ± 0.26	1.38 ± 0.00	7.12 ± 0.76
	Ignore-Include (Gaussian)	3.61 ± 0.05	6.61 ± 0.17	1.92 ± 0.15	6.90 ± 0.19
Diabetes ($N_D=10$)	Model-Analytic (Ours)	5.47 ± 0.01	45.40 ± 0.53	1.98 ± 0.07	73.40 ± 3.46
	Ignore-Include (Gaussian)	5.54 ± 0.07	48.84 ± 3.12	2.92 ± 0.14	84.89 ± 7.05
Wine Quality ($N_D=11$)	Model-Analytic (Ours)	1.38 ± 0.03	0.75 ± 0.03	0.96 ± 0.04	0.55 ± 0.04
	Ignore-Include (Gaussian)	1.18 ± 0.08	0.64 ± 0.02	2.35 ± 0.10	0.75 ± 0.03

F REAL-WORLD REGRESSION EXPERIMENTS

To assess performance in more demanding settings, we evaluate our method on a suite of real-world regression benchmarks. Our analysis considers Gaussian (‘Ignore-Include’), Truncated Gaussian (‘Ignore-Exclude’), and Censored Gaussian likelihoods. For the censored case, we explicitly compare our proposed Variational Inference against Expectation Propagation and Laplace Approximation. Extended results, covering various kernels and metrics alongside a detailed experimental protocol, are available in the Supplements.

Evaluation. We consider four benchmarks for multiple regression from continuous data: (i) Boston Housing [Harrison Jr and Rubinfeld, 1978] from OpenML, (ii) Diabetes [Efron et al., 2004] from scikit-learn and (iii) Wine Quality [Cortez et al., 2009] from UCI [Aeberhard et al., 1994]. We apply different types of censoring to each dataset. If the dataset is naturally censored we include the natural censoring and one artificial censoring to explore the potential impact. As before, we do not censor the test data to compute the latent space metrics (NLPD_f & MAE_f), but do censor the targets to compute the observation-space metrics (NLPD_c & MAE_c).

We run 5 bootstrap iterations for each method per dataset with a 50:50 train/test split. If either method fails to converge on a particular bootstrap iteration, we exclude that iteration for fairness toward both methods. For this reason, we report the Std. Err. of the mean for each metric.

Results. Table 2 shows that in most cases, Model-Analytic (Ours) is a drop-in replacement for the naive Ignore-Include (Gaussian) approach, and improves on the metrics. The Model-Analytic performance appears to suffer mildly only when the unseen tail distribution is complex—for instance, on the Wine dataset, the dependent variable represents an ordinal score ($y \in 3, 4, 5, 6, 7, 8$). Our latent metrics drop below those of the Ignore-Include (Gaussian) on that dataset due to the effect of synthetic censoring: 80% of the density is captured by the narrow uncensored interval of $[5, 7]$, which leaves no training data on the remaining half of the support.

F.1 NATURAL CENSORING ON BOSTON HOUSING

The Boston Housing dataset includes a known natural right-censoring threshold at the target value 50.. Homes with a true median value at or above \$50,000 were recorded simply as 50., which accounts for about 3% of the dataset. This creates an evaluation context where the censoring mechanism results from the data collection process itself, not from a synthetic setup. Since the true latent value beyond the threshold cannot be observed, methods can only be assessed using observed-space metrics.

In this experiment we compare three Gaussian approximations of the censored posterior: Variational Inference (CensoredGP_VI, ours), Expectation Propagation (CensoredGP_EP), and the Laplace approximation (CensoredGP_Laplace). Alongside two baselines that do not model the censoring mechanism: a standard Gaussian likelihood that treats the clamped values as clean observations (GP_Gauss), and a truncated Gaussian likelihood that drops the censored data entirely (TruncGP). We evaluate each method with both the Matérn-5/2 and RBF kernels, averaging over two folds. Results are reported in Table 3, along with wall-clock training duration.

At $\approx 3\%$ censoring, no single approximation dominates across all metrics: with the Matérn-5/2 kernel, CensoredGP_VI attains the best MAE and Hinge_MAE, CensoredGP_Laplace the best NLPD, and GP_Gauss the best RMSE, all within a narrow margin. With the RBF kernel, CensoredGP_VI is the strongest on MAE, Hinge_MAE, and RMSE,

Table 3: Predictive performance and wall-clock time on the Boston Housing dataset with the natural right-censoring threshold of 50.0 ($\approx 3\%$ of targets censored). Reported values are averaged over two folds. Lower is better for all predictive metrics; “Duration” reports total training time in seconds. Best result per column (within each kernel) is in bold.

Kernel	Method	NLPD	MAE	Hinge_MAE	RMSE	Duration (s)
Matérn-5/2	CensoredGP_EP	2.366	2.106	2.052	3.030	78.9
	CensoredGP_Laplace	2.357	2.132	2.089	3.066	27.0
	CensoredGP_VI (Ours)	2.455	2.041	2.001	3.012	56.0
	GP_Gauss	2.417	2.086	2.065	2.983	13.2
	TruncGP	2.741	2.235	2.233	3.871	12.1
RBF	CensoredGP_EP	2.403	2.300	2.210	3.316	69.9
	CensoredGP_Laplace	2.441	2.387	2.310	3.440	34.3
	CensoredGP_VI (Ours)	2.514	2.161	2.113	3.120	52.8
	GP_Gauss	2.602	2.380	2.362	3.441	13.2
	TruncGP	2.805	2.308	2.305	3.863	11.1

while `CensoredGP_EP` attains the best NLPD. This pattern is consistent with the synthetic severity study of Sec. 6.1: at low censoring fractions the systematic bias introduced by ignoring censoring is small, so the gap between the censored and uncensored likelihoods is modest, and the differences between the three Gaussian approximations for the censored likelihood are correspondingly narrow. Wall-clock training times align with the theoretical discussion of Sec. 2.2: `CensoredGP_EP` is the slowest of the three censored methods, reflecting the cost of iterative moment matching in the tails; `CensoredGP_Laplace` is the fastest, as it relies only on local curvature at the mode; and `CensoredGP_VI` sits between them, offering the best predictive accuracy of the three on this benchmark.

F.2 SYNTHETIC CENSORING ON BOSTON HOUSING

To evaluate the Gaussian approximations of the censored likelihood under controlled censoring severity, we impose synthetic upper-censoring thresholds on the Boston Housing target variable at $u = 20.0$ (censoring $\approx 58\%$ of the data) and $u = 25.0$ (censoring $\approx 26\%$). We keep the experiment setting as in Sec. F.1. Combined with the natural-censoring evaluation, this experiment gives three severity regimes ($\approx 3\%$, 26% , 58%) on the same base dataset, isolating the effect of censoring severity from all other data-specific confounders. The setting is otherwise same

We observe that the censoring severities: the gap between censored-likelihood methods and the two censoring-agnostic baselines is large and consistent: the worst censored method still improves NLPD by a factor of roughly 1.5–2.3 over the best baseline across all four task and kernel combinations. The truncated-Gaussian baseline (`TruncGP`) is uniformly the weakest, confirming that discarding censored observations is a strictly worse strategy than modelling them even when the censoring fraction is small. Second, among the three Gaussian approximations, `CensoredGP_VI` attains the best NLPD across all four task and kernel combinations and the best Hinge_MAE across all four; `CensoredGP_EP` is competitive on MAE and RMSE, edging `CensoredGP_VI` narrowly on the RBF kernel at both thresholds and on Matérn-5/2 at the milder 25.0 threshold. On wall-clock time `CensoredGP_VI` is the fastest of the three censored methods on all four combinations, running $1.5\times$ – $2\times$ faster than `CensoredGP_EP` and marginally faster than `CensoredGP_Laplace`; the uncensored baselines are naturally faster still, but at the cost of the substantial predictive degradation noted above.

F.3 ACTIVE LEARNING ON BOSTON HOUSING

To evaluate the censored BALD acquisition on a real-world regression task, we run an active learning experiment on the Boston Housing dataset with the synthetic upper-censoring threshold $u = 25.0$ ($\approx 26\%$ of targets censored). We initialize each run with $N_{\text{init}} = 10$ labelled points and grow the labelled set in increments of five up to $N = 35$, retraining the Censored GP after each acquisition step. All methods use the same underlying Censored GP surrogate with a variational Gaussian approximation of the posterior; only the acquisition score differs:

- `Rand`: uniformly random selection from the pool of unlabelled candidates.
- `BALD_gauss`: the standard Gaussian-BALD score $\frac{1}{2} \log(1 + s_x^2/\sigma_y^2)$, evaluated on the Censored GP posterior but

Table 4: Predictive performance and wall-clock training time on the Boston Housing dataset with synthetic right-censoring thresholds of 20.0 and 25.0 (approximately 58% and 26% of targets censored, respectively). Results averaged over two folds. Lower is better for all predictive metrics; “Duration” reports total training time in seconds. Best result per column (within each task–kernel block) is in bold.

Task–Kernel	Method	NLPD	MAE	Hinge_MAE	RMSE	Duration (s)
Housing_20						
Matérn-5/2	CensoredGP_EP	1.349	5.329	1.433	7.228	70.3
	CensoredGP_Laplace	1.151	3.567	1.001	5.971	49.0
	CensoredGP_VI (Ours)	1.111	3.382	0.904	5.788	41.6
	GP_Gauss	2.034	5.750	1.225	9.027	11.3
	TruncGP	2.569	7.035	2.485	10.913	2.4
Housing_20						
RBF	CensoredGP_EP	1.149	3.310	1.020	5.326	80.3
	CensoredGP_Laplace	1.174	3.456	1.049	5.661	52.7
	CensoredGP_VI (Ours)	1.121	3.466	0.914	5.837	37.2
	GP_Gauss	2.131	5.903	1.370	9.191	12.2
	TruncGP	2.430	6.684	2.119	10.263	2.3
Housing_25						
Matérn-5/2	CensoredGP_EP	1.827	2.964	1.525	4.808	57.0
	CensoredGP_Laplace	1.824	3.132	1.487	5.184	40.5
	CensoredGP_VI (Ours)	1.790	3.002	1.383	5.024	39.8
	GP_Gauss	2.146	3.913	1.450	6.827	9.7
	TruncGP	2.293	4.011	1.610	7.042	5.8
Housing_25						
RBF	CensoredGP_EP	1.860	2.999	1.557	4.810	75.5
	CensoredGP_Laplace	1.865	3.145	1.549	5.101	45.8
	CensoredGP_VI (Ours)	1.800	3.025	1.397	5.014	37.2
	GP_Gauss	2.196	3.948	1.489	6.843	10.1
	TruncGP	2.334	3.991	1.626	6.928	6.1

ignoring the censoring terms in the acquisition.

- `BALD_mc`: the exact Censored BALD score of Lemma 4.3 with the residual integral evaluated by Monte Carlo integration.
- `BALD_approx (ours)`: the closed-form Censored BALD approximation of Lemma 4.3.

Results are averaged over 5 independent runs and reported in Table 5.

Two observations. First, all three BALD variants clearly outperform random acquisition at moderate-to-large budgets: at $N = 35$, the best BALD variant improves NLPD by 5.5%, RMSE by 14.0%, and MAE by 12.6% relative to random sampling, and the separation from random becomes visible from around $N = 20$ onwards. At the smallest budget $N = 10$ all methods perform comparably, as expected since the labelled set at that budget is dominated by the initial random pool and the acquisition score has had few opportunities to influence selection. Second, the three BALD variants are close to one another at the final budget, with `BALD_gauss` attaining the best RMSE and MAE at $N = 35$, `BALD_mc` the best NLPD, and `BALD_approx (ours)` close behind on all three metrics. This demonstrates that the closed-form censored BALD approximation is a competitive replacement for the more expensive Monte Carlo estimator; the small residual gap to `BALD_gauss` on RMSE and MAE at this censoring severity (and against a Gaussian-likelihood BALD score evaluated on the same Censored GP surrogate) is consistent with the approximation-error analysis of Sec. 4: the analytic approximation is tightest in high-variance regions that greedy BALD actually targets, so the practical performance gap is small even where the score itself differs from the exact integral.

Table 5: Active learning on Boston Housing with censored regression (upper censoring threshold $u = 25.0$, $\approx 26\%$ of targets censored), averaged over 5 independent runs. All three BALD variants use the same Censored GP surrogate; only the acquisition score differs. BALD_gauss uses the standard Gaussian-BALD score $\frac{1}{2} \log(1 + s_x^2/\sigma_y^2)$, BALD_mc uses Monte Carlo integration of the exact Censored BALD score, and BALD_approx (ours) uses the closed-form approximation of Lemma 4.3. Lower is better for all metrics; best result per column in bold.

Metric	Method	$N = 10$	$N = 15$	$N = 20$	$N = 25$	$N = 30$	$N = 35$
NLPD	Rand	2.509	2.318	2.180	2.195	2.086	2.075
	BALD_mc	2.513	2.428	2.088	2.038	2.112	1.961
	BALD_gauss	2.509	2.438	2.185	2.011	1.984	1.980
	BALD_approx (Ours)	2.522	2.282	2.172	2.092	2.034	1.997
RMSE	Rand	8.226	7.343	7.010	6.723	6.450	6.264
	BALD_mc	8.228	7.962	6.310	6.092	6.496	5.836
	BALD_gauss	8.205	7.447	6.306	5.703	5.531	5.384
	BALD_approx (Ours)	8.211	7.375	6.551	6.367	6.123	5.976
MAE	Rand	5.573	4.703	4.322	4.105	3.981	3.855
	BALD_mc	5.570	5.184	3.943	3.801	4.035	3.454
	BALD_gauss	5.564	4.699	4.106	3.585	3.455	3.369
	BALD_approx (Ours)	5.554	4.630	4.218	3.979	3.745	3.571

G ADDITIONAL DETAILS

G.1 IMPACT OF CENSORING SEVERITY

All methods use a Matérn 5/2 kernel with the same initialization: kernel lengthscale $\ell = \text{std}[x]/\lambda$, kernel $\sigma_f = \text{std}[y]$, and likelihood noise $\sigma_y = \text{std}[y]/\kappa$, where $\lambda \sim U(3, 5)$ and $\kappa \sim U(5, 15)$. The variational parameters are initialized at zero with unit variance. We use a full matrix (Cholesky parameterized) covariance, and L-BFGS [Liu and Nocedal, 1989] for optimization.