
Robust Gaussian Processes via Relevance Pursuit

Sebastian Ament
Meta
ament@meta.com

Elizabeth Santorella
Meta
santorella@meta.com

David Eriksson
Meta
deriksson@meta.com

Ben Letham
Meta
bletham@meta.com

Maximilian Balandat
Meta
balandat@meta.com

Eytan Bakshy
Meta
ebakshy@meta.com

Abstract

Gaussian processes (GPs) are non-parametric probabilistic regression models that are popular due to their flexibility, data efficiency, and well-calibrated uncertainty estimates. However, standard GP models assume homoskedastic Gaussian noise, while many real-world applications are subject to non-Gaussian corruptions. Variants of GPs that are more robust to alternative noise models have been proposed, and entail significant trade-offs between accuracy and robustness, and between computational requirements and theoretical guarantees. In this work, we propose and study a GP model that achieves robustness against sparse outliers by inferring data-point-specific noise levels with a sequential selection procedure maximizing the log marginal likelihood that we refer to as *relevance pursuit*. We show, surprisingly, that the model can be parameterized such that the associated log marginal likelihood is *strongly concave* in the data-point-specific noise variances, a property rarely found in either robust regression objectives or GP marginal likelihoods. This in turn implies the weak submodularity of the corresponding subset selection problem, and thereby proves approximation guarantees for the proposed algorithm. We compare the model’s performance relative to other approaches on diverse regression and Bayesian optimization tasks, including the challenging but common setting of sparse corruptions of the labels within or close to the function range.

1 Introduction

Probabilistic models have long been a central part of machine learning, and Gaussian process (GP) models are a key workhorse for many important tasks [49], especially in the small-data regime. GPs are flexible, non-parametric predictive models known for their high data efficiency and well-calibrated uncertainty estimates, making them a popular choice for regression, uncertainty quantification, and downstream applications such as Bayesian optimization (BO) [22, 24] and active learning [6, 50].

GPs flexibly model a distribution over functions, but assume a particular observation model. The standard formulation assumes i.i.d Gaussian observation noise, i.e., $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$, where $f(\mathbf{x})$ is the true (latent) function value at a point \mathbf{x} and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, implying a homoskedastic Gaussian likelihood. While mathematically convenient, this assumption can be a limitation in practice, since noise distributions are often heavy-tailed or observations may be corrupted due to issues such as sensor failures, data processing errors, or software bugs. Using a standard GP model in such settings can result in poor predictive performance.

A number of *robust* GP modeling approaches have been proposed to remedy this shortcoming, most of which fall into the following broad categories: data pre-processing (e.g., Winsorizing), modified likelihood functions (e.g., Student- t), and model-based data selection and down-weighting procedures.

These approaches offer different trade-offs between model accuracy, degree of robustness, broad applicability, computational requirements, and theoretical guarantees.

In this paper, we propose a simple yet effective implicit data-weighting approach that endows GPs with a high degree of robustness to challenging label corruptions. Our approach is flexible and can be used with arbitrary kernels, is efficient to compute, and yields provable approximation guarantees. Our main contributions are as follows:

1. We propose a modification to the standard GP model that introduces learnable data-point-specific noise variances.
2. We introduce a novel greedy sequential selection procedure for maximizing the model’s marginal log-likelihood (MLL) that we refer to as *relevance pursuit*.
3. We prove that, under a particular parameterization, the MLL is strongly concave in the data-point-specific noise variances, and derive approximation guarantees for our algorithm.
4. We demonstrate that our approach, Robust Gaussian Processes via Relevance Pursuit (RRP), performs favorably compared to alternative methods across various benchmarks, including challenging settings of sparse label corruptions within the function’s range, see e.g. Figure 1.

2 Preliminaries

We aim to model a function $f : \mathbb{X} \rightarrow \mathbb{R}$ over some domain $\mathbb{X} \subset \mathbb{R}^d$. With a standard Gaussian noise model, for $\mathbf{x}_i \in \mathbb{X}$ we obtain observations $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. draws from a Gaussian random variable. $\|\cdot\|$ denotes the Euclidean norm unless indicated otherwise.

2.1 Gaussian Processes

A GP $f \sim \mathcal{GP}(\mu(\cdot), k_{\theta}(\cdot, \cdot))$ is fully defined by its mean function $\mu : \mathbb{X} \rightarrow \mathbb{R}$ and covariance or kernel function $k_{\theta} : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$, which is parameterized by θ . Without loss of generality, we will assume that $\mu \equiv 0$. Suppose we have collected data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{X} := \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{y} := \{y_i\}_{i=1}^n$. Let $\Sigma_{\theta} \in \mathcal{S}_{++}^n$ denote the covariance matrix of the data set, i.e., $[\Sigma_{\theta}]_{ij} = k_{\theta}(\mathbf{x}_i, \mathbf{x}_j) + \delta_{ij}\sigma^2$, where δ_{ij} is the Kronecker delta. The negative marginal log-likelihood (NMLL) \mathcal{L} is given by

$$-2\mathcal{L}(\theta) := -2\log p(\mathbf{y}|\mathbf{X}, \theta) = \mathbf{y}^{\top} \Sigma_{\theta}^{-1} \mathbf{y} + \log \det \Sigma_{\theta} + n \log 2\pi. \quad (1)$$

In the following, we will suppress the explicit dependence of the kernel matrix on θ for brevity of notation. For a comprehensive background on GPs, we refer to Rasmussen et al. [49].

2.2 Noise Models

Additive, heavy-tailed noise Instead of assuming the noise term ϵ_i in the observation model to be Gaussian, other noise models consider zero-mean perturbations drawn from distributions with heavier tails, such as the Student- t [30], Laplace [35], or α -Stable [5] distributions. These types of errors are common in applications such as finance, geophysics, and epidemiology [16]. Robust regression models utilizing Student- t errors are commonly used to combat heavy-tailed noise and outliers.

Sparse corruptions In practice, often a small number of labels are corrupted. We will refer to these as “outliers,” though emphasize that the corrupted values may fall within the range of normal outputs. Sparse corruptions are captured by a model of the form $y_i = Z_i f(\mathbf{x}_i) + (1 - Z_i)W_i$, where

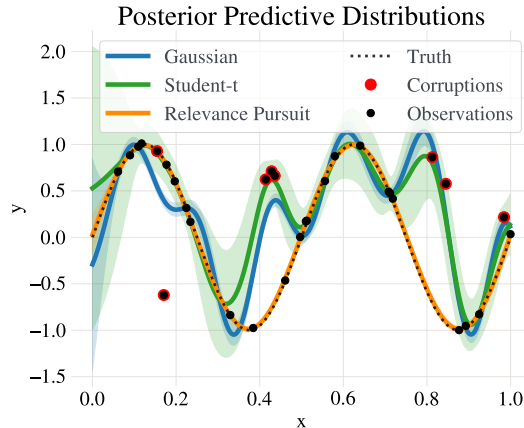


Figure 1: Comparison of RRP to a standard GP and a variational GP with a Student- t likelihood on a regression example. While the other models are led astray by the corrupted observations, RRP successfully identifies the corruptions (red) and thus achieves a much better fit to the ground truth.

$Z_i \in \{0, 1\}$ and $W_i \in \mathbb{R}$ is a random variable. Note that W_i need not have (and rarely has) $f(\mathbf{x}_i)$ as its mean. For instance, consider a faulty sensor that with some probability p reports a random value within the sensor range $[y_l, y_h]$. In this case $Z_i \sim \text{Ber}(p)$ and $W_i \sim U[y_l, y_h]$. Software bugs, such as those found in ML training procedures, or errors in logging data can result in sparse corruptions.

3 Related Work

Data pre-processing Data pre-processing can be an effective technique for handling simple forms of data corruption, such as values that fall outside a valid range of outputs. With such pre-processing, outliers are handled upstream of the regression model. Common techniques include the power transformations [13], trimming, and winsorization. These methods can add substantial bias if not used carefully, and generally do not handle data corruptions that occur within the normal range of the process to be modeled. See [15] for a review on data cleaning.

Heavy-tailed likelihoods One class of robust methods uses additive heavy-tailed noise likelihoods for GPs, particularly Student- t [30], Laplace [48], and Huber [1], and could be extended with α -Stable distributions, which follow a generalized central limit theorem [5]. These models are less sensitive to outliers, but they lose efficiency when the outliers are a sparse subset of the observations, as opposed to global heavy-tailed noise. Furthermore, model inference is no longer analytic, necessitating the use of approximate inference approaches such as MCMC [44], Laplace approximation [59], expectation propagation (EP) [30], Expectation Maximization [48], or variational inference [56]. Shah et al. [51] take a related approach using a Student- t process prior in the place of the GP prior. Unfortunately, the Student- t process is not closed under addition and lacks the tractability that makes GPs so versatile. Alternative noise specifications include a hierarchical mixture of Gaussians [17] and a “twinned” GP model [43] that uses a two-component noise model to allow outlier behavior to depend on the inputs. This method is suited for settings where outliers are not totally stochastic, but generally is not able to differentiate “inliers” from outliers when they can occur with similar inputs.

Outlier classification Awasthi et al. [10] introduces the Trimmed MLE approach, which identifies the subset of data points (of pre-specified size) under which the marginal likelihood is maximized. Andrade and Takeda [9] fit GPs using the trimmed MLE by applying a projected gradient method to an approximation of the marginal likelihood. The associated theory only guarantees convergence to a stationary point, with no guarantee on quality. When no outliers are present, this method can be worse than a standard GP. Li et al. [36] propose a heuristic iterative procedure of removing those data points with the largest residuals after fitting a standard GP, with subsequent reweighting. The method shows favorable empirical performance but has no theoretical guarantees, and fails if the largest residual is not associated with an outlier. Park et al. [47] consider a model of the form $y_i = \delta_i + f(\mathbf{x}_i) + \epsilon_i$, where outliers are regarded as data with a large bias δ_i . Their random bias model is related to our model in that it also introduces learnable, data-point-specific variances. However, inference is done in one step by optimizing the NMLL with an inverse-gamma prior on the δ_i 's, which – in contrast to the method proposed herein – generally does not lead to exactly sparse δ_i 's.

Sample re-weighting Altamirano et al. [2] propose robust and conjugate GPs (RCGP) based on a modification to the Gaussian likelihood function that is equivalent to standard GP inference, where the covariance of the noise $\sigma^2 \mathbf{I}$ is replaced by $\sigma^2 \text{diag}(\mathbf{w}^{-2})$ and the prior mean \mathbf{m} is replaced by $\mathbf{m}_w = \mathbf{m} + \sigma^2 \nabla_y \log(\mathbf{w}^2)$. The authors advocate for the use of the inverse multi-quadratic weight function $w(\mathbf{x}, y) = \beta(1 + (y - m(\mathbf{x}))^2/c^2)^{-1/2}$, which introduces two additional hyper-parameters: the soft threshold c , and the “learning rate” β . Importantly, the weights \mathbf{w} are defined *a-priori* as a function of the prior mean $m(\mathbf{x})$ and the targets y , thereby necessitating the weights to identify the correct outliers without access to a model. This is generally only realistic if the outlier data points are clearly separated in the input or output spaces rather than randomly interspersed.

4 Robust Gaussian Process Regression via Relevance Pursuit

Our method adaptively identifies a sparse set of outlying data points that are corrupted by a mechanism that is not captured by the other components of the model. This is in contrast to many other approaches to robust regression that non-adaptively apply a heavy-tailed likelihood to *all* observations, which can be suboptimal if many observations are of high quality.

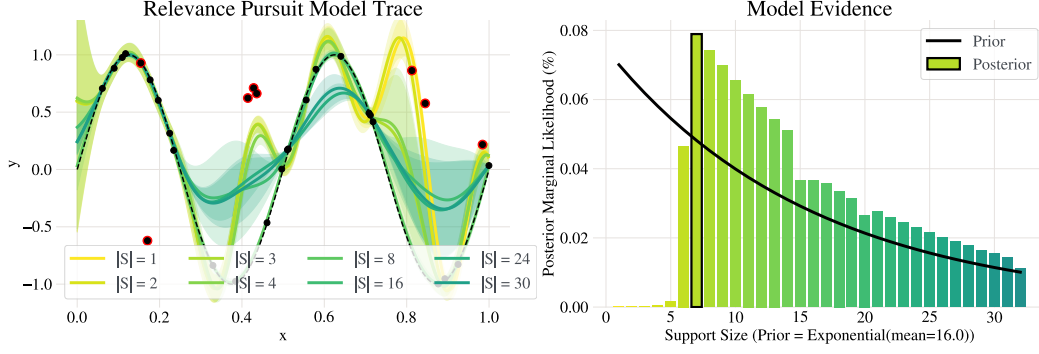


Figure 2: *Left*: Evolution of model posterior during Relevance Pursuit, as the number of data-point-specific variances $|S|$ increases (from light colors to dark). Red points indicate corruptions that were generated by uniformly sampling from the function’s range. *Right*: Comparison of posterior marginal likelihoods as a function of a model’s $|S|$. The maximizer – boxed in black – is the preferred model.

4.1 The Extended Likelihood Model

We extend the standard GP observation noise variance σ^2 with data-point-specific noise variances $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^n$, so that the i -th data point is distributed as

$$y_i \mid \mathbf{x}_i \sim \mathcal{N}(f(\mathbf{x}_i), \sigma^2 + \rho_i). \quad (2)$$

This is similar to Sparse Bayesian Learning [55] in which weight-specific prior variances control a feature’s degree of influence on a model’s predictions. The likelihood in (2) is similar in that the marginal likelihood optimization of the ρ_i gives rise to an *automatic mechanism* for outlier detection and down-weighting. Intuitively, the effect of y_i on the estimate of f vanishes as $\rho_i \rightarrow \infty$. This model can be viewed as a type of GP with heteroskedastic noise model. However, while many heteroskedastic models model noise as an input-dependent process [26, 33], our formulation does not make such assumptions, and is thus suitable for corruptions that are not spatially correlated.

An elegant consequence of our modeling assumption is that we can compute individual marginal-likelihood maximizing ρ_i ’s in closed form when keeping all ρ_j for $j \neq i$ fixed. In particular,

Lemma 1. [Optimal Robust Variances] Let $\mathcal{D}_{\setminus i} = \{(\mathbf{x}_j, y_j) : j \neq i\}$, $\boldsymbol{\rho} = \boldsymbol{\rho}_{\setminus i} + \rho_i \mathbf{e}_i$, where $\boldsymbol{\rho}, \boldsymbol{\rho}_{\setminus i} \in \mathbb{R}_+^n$, $[\boldsymbol{\rho}_{\setminus i}]_i = 0$, and \mathbf{e}_i is the i th canonical basis vector. Then keeping $\boldsymbol{\rho}_{\setminus i}$ fixed,

$$\rho_i^* = \arg \max_{\rho_i} \mathcal{L}(\boldsymbol{\rho}_{\setminus i} + \rho_i \mathbf{e}_i) = [(y_i - \mathbb{E}[y(\mathbf{x}_i) \mid \mathcal{D}_{\setminus i}])^2 - \mathbb{V}[y(\mathbf{x}_i) \mid \mathcal{D}_{\setminus i}]]_+, \quad (3)$$

where $y(\mathbf{x}_i) = f(\mathbf{x}_i) + \epsilon_i$. These quantities can be expressed as functions of $\boldsymbol{\Sigma}^{-1} = (\mathbf{K} + \mathbf{D}_{\sigma^2 + \boldsymbol{\rho}})^{-1}$:

$$\mathbb{E}[y(\mathbf{x}_i) \mid \mathcal{D}_{\setminus i}]^2 = y_i - [\boldsymbol{\Sigma}^{-1} \mathbf{y}]_i / [\boldsymbol{\Sigma}^{-1}]_{ii}, \quad \text{and} \quad \mathbb{V}[y(\mathbf{x}_i) \mid \mathcal{D}_{\setminus i}] = 1 / [\boldsymbol{\Sigma}^{-1}]_{ii},$$

where $\mathbf{D}_{\sigma^2 + \boldsymbol{\rho}}$ is a diagonal matrix whose entries are $\sigma^2 + \rho$.

The first component $\mathbb{E}[f(\mathbf{x}_i) + \epsilon_i \mid \mathcal{D}_{\setminus i}]^2$ of (3) is the empirical error to y_i of the model trained without the i -th data point, i.e., the leave-one-out (LOO) cross-validation error [49]. The second component $\mathbb{V}[f(\mathbf{x}_i) + \epsilon_i \mid \mathcal{D}_{\setminus i}]$ is the LOO predictive variance. The optimal solution to ρ_i is only non-zero for those observations whose squared LOO error is larger than the LOO predictive variance at that point.

4.2 Optimization with a Maximum Number of Outliers

Without additional structure, inference of the noise variances ρ_i does not yield desirable models, as the marginal likelihood can be improved by increasing the prior variance ρ_i of any data point where Eq. (3) is greater than zero, even if that is due to regular (non-outlier) measurement noise. To avoid this, we constrain the number of non-zero ρ_i , that is, $\|\boldsymbol{\rho}\|_\infty = |\{0 < \rho_i\}| \leq k < n$. While this sparsity constraint mitigates over-flexibility, it gives rise to a formidably challenging optimization problem, as there are a combinatorial number of sparse outlier sets to consider. Even if the number of outliers n_o were known, exhaustive search would still require considering n -choose- n_o possibilities.

For tractability, we iteratively add data points to a set of potential “outliers” by setting their associated ρ_i to be nonzero, using the closed-form expression for the optimal individual ρ_i variances in Lemma 1. As the algorithm seeks to identify the most “relevant” data points (as measured by \mathcal{L}) upon completion, we refer to it as *Relevance Pursuit*. This is Algorithm 1 with `useBayesianModelSelection` as `false`. Specifically, this is the “forward” variant; Algorithm 2 in the Appendix presents an alternative “backward” variant that we found to work well if the number of corrupted data points is large.

Crucial to the performance of the optimizer, it never removes data from consideration completely; a data point is only down-weighted if it is apparently an outlier. This allows the down-weighting to be reversed if a data point appears “inlying” after having down-weighted other data points, improving the method’s robustness and performance. This is in contrast to Andrade and Takeda [9]’s greedy algorithm, in which the exclusion of a data point can both increase or decrease the associated marginal likelihood. This means that their objective is not monotonic, a necessary condition to provide constant-factor submodular approximation guarantees for greedy algorithms, see Section 5.

Algorithm 1 Relevance Pursuit (Forward Algorithm)

Require: \mathbf{X}, \mathbf{y} , schedule $\mathcal{K} = (k_1, k_2, \dots, k_{|\mathcal{K}|})$, `useBayesianModelSelection` (boolean)

Initialize $\mathcal{S}_0 \subseteq \{1, \dots, n\}$ (typically $\mathcal{S}_0 = \emptyset$)

for i in $(1, \dots, |\mathcal{K}|)$ **do**

Optimize MLL: $\rho_{\mathcal{S}_i} \leftarrow \arg \max_{\rho_{\mathcal{S}_i}} \mathcal{L}(\rho_{\mathcal{S}_i})$, where $\rho_{\mathcal{S}_i} = \{\rho : \rho_j = 0, \forall j \notin \mathcal{S}_i\}$.

Expand Support:

Compute $\Delta_i(j) \leftarrow \max_{\rho_j} \mathcal{L}(\rho_{\mathcal{S}_i} + \rho_j \mathbf{e}_j) - \mathcal{L}(\rho_{\mathcal{S}_i})$ for all $j \notin \mathcal{S}_i$ via Lemma 1.

$\mathcal{A}_i \leftarrow \{j_1, \dots, j_{k_i}\}$ such that $\Delta_i(j) \geq \Delta_i(j')$ for all $j \in \mathcal{A}_i$ and $j' \notin (\mathcal{A}_i \cup \mathcal{S}_i)$.

$\mathcal{S}_{i+1} \leftarrow \mathcal{S}_i \cup \mathcal{A}_i$

if `useBayesianModelSelection` **then**

Compute the marginal likelihood $p(\mathcal{D}|\mathcal{S}_i) \approx p(\mathcal{D}|\mathcal{S}_i, \rho_{\mathcal{S}_i})$

$\mathcal{S}^* \leftarrow \arg \max_{\mathcal{S}_i} p(\mathcal{D}|\mathcal{S}_i)p(\mathcal{S}_i)$.

else

$\mathcal{S}^* = \mathcal{S}_{|\mathcal{K}|}$.

Return $\mathcal{S}^*, \rho_{\mathcal{S}^*}$.

4.3 Automatic Outlier Detection via Bayesian Model Selection

In practice, it is often impossible to set a hard threshold on the number of outliers for a particular data set. For example, a sensor might have a known failure rate, but how many outliers it produces will depend on the specific application of the sensor. Thus, it is often more natural to specify a prior distribution $p(\mathcal{S})$ over the number of outliers, rather than fix the number *a priori*. We leverage the Bayesian model selection framework [61, 40] to determine the most probable number of outliers in a data- and model-dependent way, aiming to maximize $p(\mathcal{S}|\mathcal{D})$. This gives rise to Algorithm 1, with `useBayesianModelSelection` as `true`.

Computationally, we start by iteratively adding outliers up to the maximal support of the prior, similar to the procedure described in Section 4.2. We store a trace of models generated at each iteration, then approximate the model posterior $p(\mathcal{S}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{S}_i)p(\mathcal{S}_i)$ at each point in the trace. As the exact posterior is intractable, we approximate it with $p(\mathcal{D}|\mathcal{S}_i) = \int p(\mathcal{D}|\mathcal{S}_i, \rho_{\mathcal{S}_i})d\rho_{\mathcal{S}_i} \approx p(\mathcal{D}|\mathcal{S}_i, \rho_{\mathcal{S}_i}^*)$. Finally, we select the model from the model trace $\{\mathcal{S}_i\}_i$ that attains the highest model posterior likelihood. Imposing a prior on the number of outliers differs notably from most sparsity-inducing priors, which are instead defined on the parameter values, like l_1 -norm regularization. In practice, $p(\mathcal{S})$ can be informed by empirical distributions of outliers. For our experiments, we use an exponential prior on $|\mathcal{S}|$ to encourage the selection of models that fit as much of the data as tightly as possible.

Regarding the schedule \mathcal{K} in Algorithm 1, the most natural choice is simply to add one data point at a time, i.e. $\mathcal{K} = (1, 1, \dots)$, but this can be slow for large n . In practice, we recommend schedules that test a fixed set of outlier fractions, e.g. $\mathcal{K} = (0.05n, 0.05n, \dots)$.

5 Theoretical Analysis

We now provide a theoretical analysis of our approach. We first propose a re-parameterization of the ρ_i that maps the optimization problem to a compact domain. Surprisingly, the re-parameterized

problem exhibits strong convexity and smoothness when the base covariance matrix (excluding the ρ_i) is well-conditioned. We connect the convexity and smoothness with existing results that yield approximation guarantees for sequential greedy algorithms, implying a constant-factor approximation guarantee to the optimal achievable NMLL value for generalized orthogonal matching pursuit (OMP), a greedy algorithm that is closely related to Algorithm 1.

5.1 Preliminaries for Sparse Optimization

The optimization of linear models with respect to least-squares objectives in the presence of sparsity constraints has been richly studied in statistics [54], compressed sensing [3, 57], and machine learning [8, 62]. Of central importance to the theoretical study of this problem class are the eigenvalues of sub-matrices of the feature matrix, corresponding to sparse feature selections and so often referred to as *sparse eigenvalues*. The restricted isometry property (RIP) formalizes this.

Definition 2 (Restricted Isometry Property). *An $(n \times m)$ -matrix \mathbf{A} satisfies the r -restricted isometry property (RIP) with constant $\delta_r \in (0, 1)$ if for every submatrix $\mathbf{A}_{\mathcal{S}}$ with $|\mathcal{S}| = r \leq m$ columns,*

$$(1 - \delta_r)\|\mathbf{x}\| \leq \|\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\| \leq (1 + \delta_r)\|\mathbf{x}\|,$$

where $\mathbf{x}_{\mathcal{S}} \in \mathbb{R}^r$. This is equivalent to $(1 - \delta_r)\mathbf{I} \preceq (\mathbf{A}_{\mathcal{S}}^*\mathbf{A}_{\mathcal{S}}) \preceq (1 + \delta_r)\mathbf{I}$.

The RIP has been proven to lead to exact recovery guarantees [14], as well as approximation guarantees [18]. Elenberg et al. [21] generalized the RIP to non-linear models and other data likelihoods, using the notion of restricted strong convexity (RSC) and restricted smoothness.

Definition 3 (Restricted Strong Convexity and Smoothness). *A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is m_r -restricted strong convex and M_r -restricted smooth if for all $(\mathbf{x}, \mathbf{x}')$ in the domain $D_r \subset (\mathbb{R}^d \times \mathbb{R}^d)$,*

$$m_r\|\mathbf{x}' - \mathbf{x}\|^2/2 \leq f(\mathbf{x}') - f(\mathbf{x}) - \nabla[f](\mathbf{x})^\top(\mathbf{x}' - \mathbf{x}) \leq M_r\|\mathbf{x}' - \mathbf{x}\|^2/2.$$

In the context of sparse optimization, we let D_r be the set of tuples of r -sparse vectors whose difference is also at most r -sparse. In particular, $D_r = \{(\mathbf{x}, \mathbf{x}') \text{ s.t. } \|\mathbf{x}\|_0, \|\mathbf{x}'\|_0, \|\mathbf{x}' - \mathbf{x}\|_0 \leq r\}$.

Generalized orthogonal matching pursuit (OMP) [4, 39] is a greedy algorithm that keeps track of a support set \mathcal{S} of non-zero coefficients, and expands the support based on the largest gradient magnitudes, applied to the marginal likelihood optimization problem, $\mathcal{S}_{i+1} = \mathcal{S}_i \cup \arg \max_{j \notin \mathcal{S}} |\nabla_{\rho} \mathcal{L}(\boldsymbol{\rho})|_j$. Algorithm 1 generalizes OMP [57] by allowing more general support expansion schedules \mathcal{K} , and specializes the support expansion criterion using the special problem structure exposed by Lemma 1.

5.2 The Convex Parameterization

The NMLL \mathcal{L} of a GP (1) is the sum of a convex function $(\cdot)^{-1}$ and a concave function $\log \det(\cdot)$ of \mathbf{K} , and is therefore not generally convex as a function of the hyper-parameters $\boldsymbol{\theta}$, including the robust variances $\boldsymbol{\rho}$. Here, we propose a re-parameterization that allows us to prove strong convexity guarantees of the associated NMLL. In particular, we let $\boldsymbol{\rho}(\mathbf{s}) = \text{diag}(\mathbf{K}_0) \odot ((1 - \mathbf{s})^{-1} - 1)$, where $\mathbf{K}_0 := k(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}$ and the inverse is element-wise. Note that $\boldsymbol{\rho}(\mathbf{s})$ is a diffeomorphism that maps \mathbf{s} from the compact domain $\mathbf{s} \in [0, 1]^n$ to the entire range of $\boldsymbol{\rho} \in [0, \infty]^n$.

Henceforth, we refer to the original $\boldsymbol{\rho}$ as the *canonical* or $\boldsymbol{\rho}$ -parameterization and the newly proposed $\boldsymbol{\rho}(\mathbf{s})$ as the *convex* or \mathbf{s} -parameterization. Lemma 4 shows the Hessian of the \mathbf{s} -parameterization.

Lemma 4. [Reparameterized Hessian] *Let $\mathbf{K}_{\mathbf{s}} = k(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I} + \mathbf{D}_{\boldsymbol{\rho}(\mathbf{s})}$, $\hat{\mathbf{K}}_{\mathbf{s}} = \text{diag}(\mathbf{K}_{\mathbf{s}})^{-1/2}\mathbf{K}_{\mathbf{s}}\text{diag}(\mathbf{K}_{\mathbf{s}})^{-1/2}$, and $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{K}}_{\mathbf{s}}^{-1}\text{diag}(\mathbf{K}_{\mathbf{s}})^{-1/2}\mathbf{y}$. Then*

$$\mathbf{H}_{\mathbf{s}}[-2\mathcal{L}(\boldsymbol{\rho}(\mathbf{s}))] = \mathbf{D}_{1-\mathbf{s}}^{-1} \left[2 \left(\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) + 2 \text{diag}(\hat{\mathbf{K}}^{-1}) - (\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}) \right] \mathbf{D}_{1-\mathbf{s}}^{-1}.$$

Based on this representation, we now derive conditions on the eigenvalues of $\hat{\mathbf{K}}$ that imply the m -strong convexity and M -smoothness of the NMLL.

Lemma 5. [Strong Convexity via Eigenvalue Condition] *Let $\hat{\mathbf{K}}_{\mathbf{s}}$ as in Lemma 4. Then $\mathbf{H}_{\mathbf{s}} \succ m$ if*

$$\lambda_{\min}\hat{\lambda}_{\min}^2 \frac{(2\hat{\lambda}_{\max}^{-1} - \hat{\lambda}_{\min}^{-2} - m)}{2(1 - \lambda_{\min}/\lambda_{\max})} > \|\mathbf{y}\|_2^2, \quad (4)$$

where $\lambda_{\min, \max}$ (resp. $\hat{\lambda}_{\min, \max}$) are the smallest and largest eigenvalues of $\mathbf{K}_{\mathbf{s}}$, respectively $\hat{\mathbf{K}}_{\mathbf{s}}$.

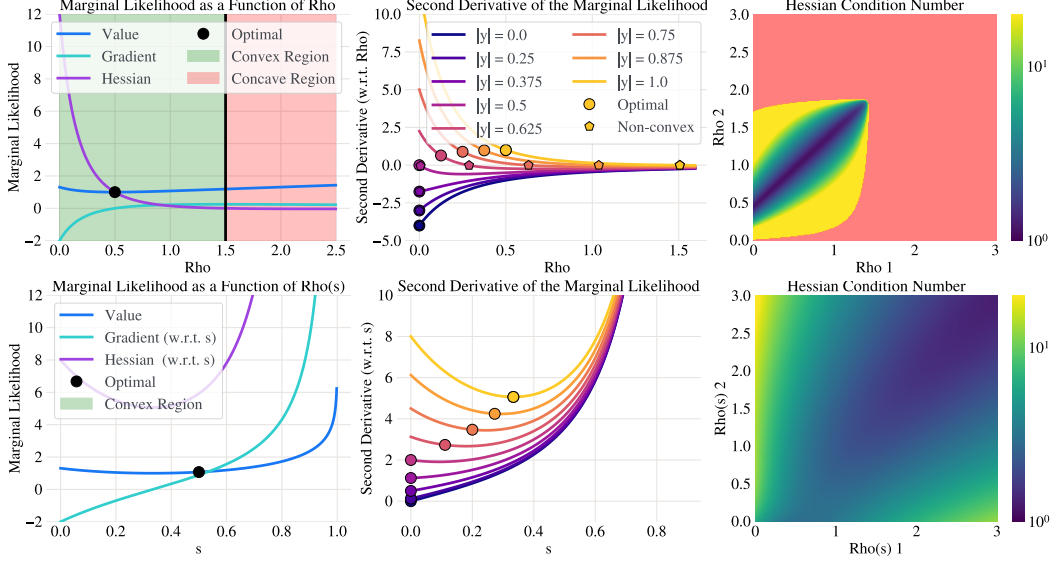


Figure 3: *Top:* The behavior of the $-\log \mathcal{L}(\rho)$ with respect to the canonical parameterization of ρ . *Bottom:* The behavior of $-\log \mathcal{L}(\rho(s))$, highlighting the convexity property. *Left:* The value, and first two derivatives of $-\log \mathcal{L}$ for a 1d example. *Center:* The second derivatives of a 1d $-\log \mathcal{L}$ as a function of $|y|$. The s -parameterization is everywhere convex for all considered $|y|$, while the canonical ρ -parameterization is only convex around the origin and only for $|y| > 0.5$. *Right:* The heatmaps highlight that the original parameterization is non-convex (red) for larger values of ρ , and quickly becomes ill-conditioned, whereas the parameterization $\rho(s)$ is convex and much better conditioned.

The behavior Lemma 5 predicts is surprising and validated in Fig. 3. Notably, the denominator “blows up” as \mathbf{K} becomes close to unitary, making the inequality more likely to be satisfied, an indication that the convexity property of the NMLL is intimately linked to the RIP (Def. 2). Note that Lemma 5 is a condition for non-support-restricted convexity, which is stronger than is necessary for the approximation guarantees that rely on restricted convexity (Def. 3). However, sparse eigenvalues are generally difficult to compute exactly. Fortunately, covariance matrices of GPs naturally tend to exhibit a property that facilitates a different sufficient condition for convexity for all $s \in [0, 1]^n$.

Definition 6 (Diagonal Dominance). *A matrix \mathbf{A} is said to be δ -diagonally dominant if the elements a_{ij} satisfy $\sum_{i \neq j} |a_{ij}| < \delta |a_{ii}|$ for all i .*

Intuitively, the $\rho_i(s)$ that are selected to be non-zero by the greedy algorithm take on large values, further encouraging the diagonal dominance of the sub-matrix of \mathbf{K} associated with the support of ρ . For this reason, the following condition on \mathbf{K}_0 is sufficient to guarantee convexity for all $s \in [0, 1]^n$.

Lemma 7. [Strong Convexity via Diagonal Dominance] *Let $m > 0$ and \mathbf{K}_0 be δ -diagonally dominant with $\delta < ((5 - m) - \sqrt{25 - 9m + 17}) / 4 \leq (5 - \sqrt{17}) / 4 \approx 0.44$ and*

$$\lambda_{\min}(\mathbf{K}_0)(1 - \delta)^2 \frac{2(1 + \delta)^{-1} - (1 - \delta)^{-2} - m}{2(1 - (1 - \delta)/(1 + \delta))} \geq \|\mathbf{y}\|_2^2.$$

Then the NMLL is m -strongly convex for all $s \in [0, 1]^n$, i.e. $\rho(s) \in [0, \infty]^n$.

We attain similar results for M -smoothness, see Lemma 13 and Lemma 14 in the Appendix. Having proven m -convexity and M -smoothness conditions, we appeal to the results of Elenberg et al. [21].

Theorem 8. [Approximation Guarantee] *Let $\mathbf{K}_0 = k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}$ be δ -diagonally dominant, $s_{\max} > 0$ be an upper bound on $\|\mathbf{s}\|_{\infty}$, and suppose $\|\mathbf{y}\|, \delta$ satisfy the bounds of Lemmas 7 and 14, guaranteeing m -convexity and M -smoothness of the NMLL for some $m > 0, M > 1/(1 - s_{\max})^2$. Let $\mathbf{s}_{\text{OMP}}(r)$ be the r -sparse vector attained by OMP on the NMLL objective for r steps, and let $\mathbf{s}_{\text{OPT}}(r) = \arg \max_{\|\mathbf{s}\|_0=r, \|\mathbf{s}\|_{\infty} \leq s_{\max}} \mathcal{L}(\rho(\mathbf{s}))$ be the optimal r -sparse vector. Then for any $2r \leq n$,*

$$\tilde{\mathcal{L}}(\rho(\mathbf{s}_{\text{OMP}}(r))) \geq (1 - e^{-m/M}) \tilde{\mathcal{L}}(\rho(\mathbf{s}_{\text{OPT}}(r))),$$

where $\tilde{\mathcal{L}}(\cdot) = \mathcal{L}(\cdot) - \mathcal{L}(\mathbf{0})$ is normalized so that $\max_{\mathbf{s}_S} \tilde{\mathcal{L}}(\mathbf{s}_S) \geq 0$ for any support S .

A limitation of the theory is that it assumes the other hyper-parameters of the GP model to be constant, as doing otherwise would introduce the non-convexity that is common to most marginal likelihood optimization problems. In practice, we typically optimize ρ jointly with the other hyper-parameters of the model in each iteration of RRP, as this yields improved performance, see App. D.5 for details.

6 Empirical Results

We evaluate the empirical performance of RRP against various baselines on a number of regression and Bayesian Optimization problems. Specifically, we compare against a standard GP with a Matern-5/2 kernel (“Standard GP”), data pre-processing through Ax’s adaptive winsorization procedure (“Adapt. Wins.”) [11], and a power transformation (“Power Transf.”) [13]. Further, we also consider a Student- t likelihood model from Jylänki et al. [30] (“Student- t ”), the trimmed marginal likelihood model from Andrade and Takeda [9] (“Trimmed MLL”), and the RCGP model from Altamirano et al. [2]. Unless stated otherwise, all models are implemented in GPyTorch [23] and all experiments in this section use 32 replications. See Appendix D for additional details.

6.1 Regression Problems

Synthetic We first consider the popular Friedman10 and Hartmann6 [19] test functions from the literature. We use two data generating processes: uniform noise, extreme outliers at some fixed value, and heavy-tailed (Student- t) noise at true function values. In these experiments, we compare the performance predictive log-likelihood. The results are shown in Fig. 4.

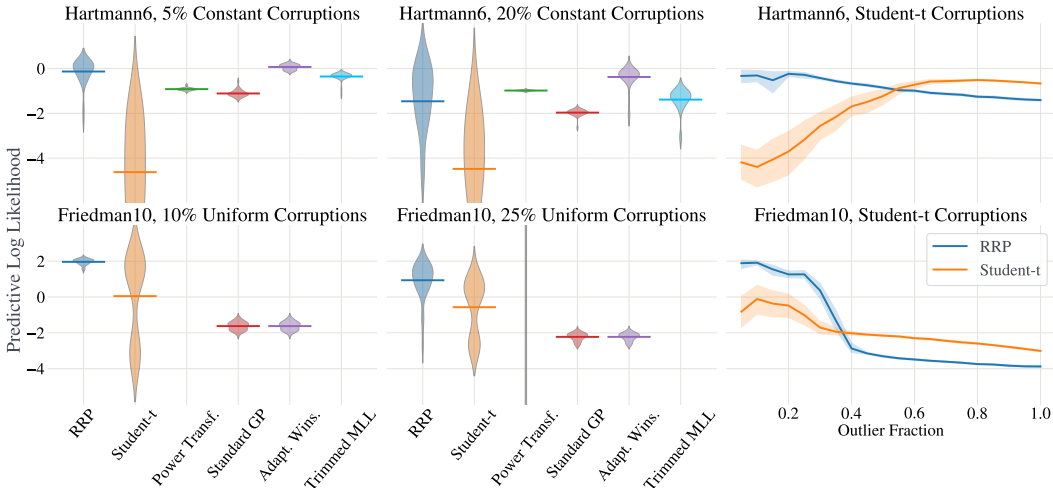


Figure 4: *Left*: Distribution of predictive test-set log likelihood for various methods. Methods omitted are those that performed substantially worse. *Right*: Predictive log likelihood as a function of the corruption probability for Student- t -distributed corruptions with two degrees of freedom. The GP model with the Student- t likelihood only starts outperforming RRP as the corruption probability increases beyond 40%, and exhibits a large variance in outcomes, which shrinks as the proportion of corruptions increases. All methods not shown were inferior to either RRP or Student- t .

Twitter Flash Crash In Fig. 5, we report a comparison to Altamirano et al. [2]’s RCGP on data from the Dow Jones Industrial Average (DJIA) index on April 22-23 2013, which includes a sharp drop at 13:10 on the 23rd. The top panels shows that RCGP exhibits higher robustness than the standard GP, but is still affected by the outliers, when trained on data from the 23rd. RRP is virtually unaffected. Notably, RCGP relies on an a-priori weighting of data points based on the target values’ proximity to their median, which can be counter-productive when the outliers are not a-priori separated in the range. To highlight this, we included the previous trading day into the training data for the bottom panels, leading RCGP to assign the *highest* weight to the outlying data points due to their proximity to the target values’ median, thereby leading RCGP to “trust” the outliers more than any inlier,

resulting in it being less robust than a standard GP in this scenario. See Appendix D.6 for additional comparisons to RCGP, on data sets from the UCI machine learning repository [32].

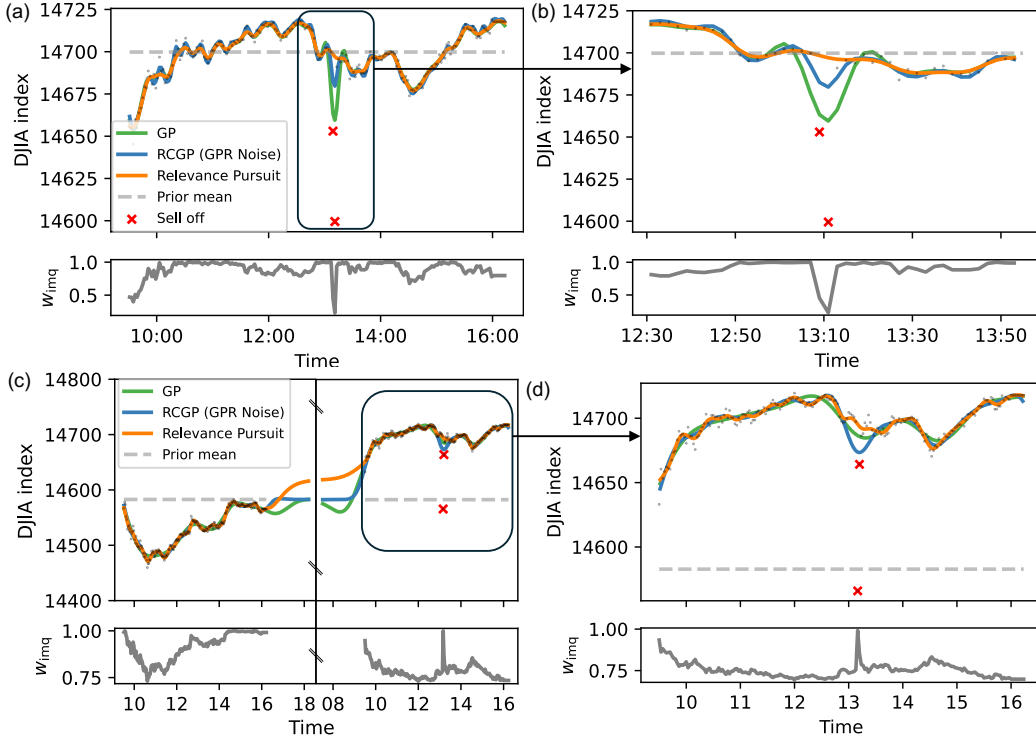


Figure 5: Results on the intra-day data from the Dow Jones Industrial Average (DJIA) index on April 22-23 2013, which includes a sharp drop at 13:10 on the 23rd, see (b) for a detailed view. The accompanying panels labeled w_{imq} show the function that Altamirano et al. [2]’s RCGP uses to down-weight data points. *Top*: RCGP, exhibits higher robustness than the standard GP, but is still affected by the outliers. The RRP model is virtually unaffected. *Bottom*: Including the previous trading day into the training data in (c), leads RCGP to assign the *highest* weight w_{imq} to the outlying data points due to their proximity to the target values’ median, thereby leading RCGP to be even more affected than a standard GP, see (d) for a detailed view of the results on the data of April 23.

6.2 Robust Bayesian Optimization

GPs are commonly used for Bayesian optimization (BO), which is a popular approach to sample-efficient black-box optimization [22]. However, many of the GP models used for BO are sensitive to outliers and may not perform well in settings where such outliers occur. While Martinez-Cantin et al. [41] consider the use of a Student- t likelihood for BO with outliers, the use of other robust GP models has not been thoroughly studied in the literature.

Experimental setup We use Ament et al. [7]’s qLogNoisyExpectedImprovement (qLogNEI), a variant of the LogEI family of acquisition functions, 32 replications, and initialize all methods with the same quasi-random Sobol batch for each replication. We follow Hvarfner et al. [29] and plot the true value of the best in-sample point according to the GP model posterior at each iteration. We also include Sobol and an “Oracle”, which is a Standard GP that always observes the uncorrupted value, and consider the backward canonical version of relevance pursuit, denoted by RRP, for these experiments. The plots show the mean performance with a bootstrapped 90% confidence interval.

Synthetic problems We consider the popular 6-dimensional Hartmann test function with three different corruption settings: (1) a constant value of 100, (2) a $U[-3, 3]$ distributed value, (3) the objective value for a randomly chosen point in the domain. The results for a 10% corruption probability are shown in Fig. 6. We also include results for a 20% corruption probability in Appendix D.3.

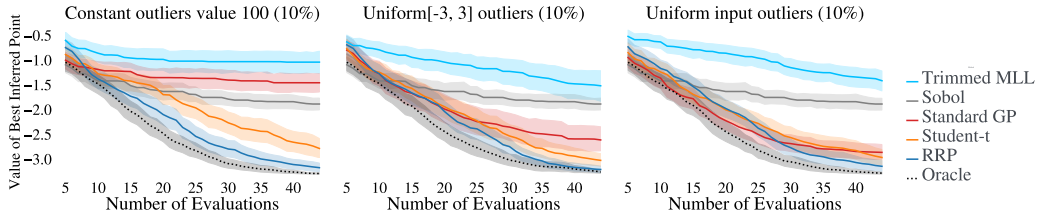


Figure 6: BO results for Hartmann6: *Left*: Relevance pursuit performs well in the case of constant outliers of value 100, almost as well as the oracle. *Middle*: Relevance pursuit performs the best followed by the Student- t likelihood in the case of $U[-3, 3]$. *Right*: Similar to the middle plot, this setting hides the corruptions within the range of the function, making it a challenging task.

Real-world problems We include three real-world problems: A 3D SVM problem, a 5D CNN problem, and a 20D rover trajectory planning problem, see the App. D.2 for details. For SVM and CNN, we simulate random corruptions corresponding to an I/O error, which causes the corresponding ML model to be trained using only a small subset of the training data. For the rover planning problem we follow the setup in [42] with the main difference that we consider a 20D trajectory, and the corruptions are generated randomly, causing the rover to break down at an arbitrary point along its trajectory. In most cases, this results in a smaller reward than the reward of the full trajectory.

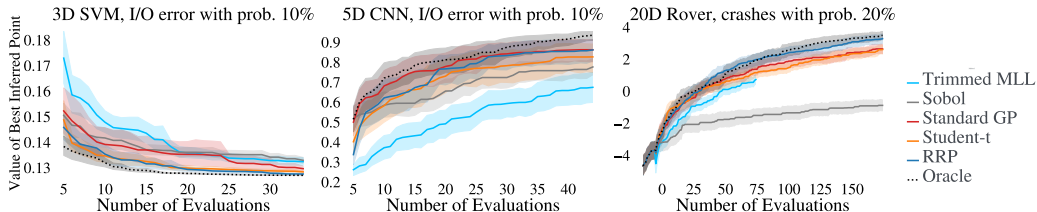


Figure 7: BO results for three real-world problems: *Left*: RRP is competitive with the oracle on the 3D SVM problem. *Middle*: The power transform performs best on the 5D CNN problem, outperforming RRP as well as the Oracle. *Right*: RRP performs well on the 20D Rover problem.

7 Conclusion and Future Work

Contributions Robust Gaussian Processes via Relevance Pursuit (RRP) provides a novel and principled way to perform robust GP regression. It permits efficient and robust inference, performs well across a variety of label corruption settings, retains good performance in the absence of corruptions, and is flexible, e.g., can be used with any mean or kernel function. Our method can be readily applied to both robust regression problems as well as applications such as Bayesian optimization and is available through BoTorch [12]. Importantly, it also provides theoretical approximation guarantees.

Limitations As our approach does not explicitly consider the locations of the data points in the outlier identification, it may be outperformed by other methods if the underlying noise is heteroskedastic and location-dependent. On the other hand, those methods generally do not perform well in the presence of sparse, location-independent data corruptions.

Extensions Promising extensions of this work include performing Bayesian model averaging, i.e., average the predictions of the different possible sparsity models according to their likelihoods instead of using a MAP estimate, applying RRP to specialized models such as Lin et al. [37]’s scalable learning-curve model for AutoML applications, and leveraging forward-backward greedy algorithms. On a higher level, the approach of combining greedy optimization algorithms with Bayesian model selection and leveraging a convex parameterization to achieve approximation guarantees might apply to other parameters that are optimized using the MLL objective: length-scales of stationary kernels, coefficients of additive kernels, inducing inputs, and even related model classes like Tipping [55]’s Sparse Bayesian Learning (SBL), which seeks to identify sparse linear models and is intimately linked to greedy matching pursuits [8]. Overall, the approach has the potential to lead to theoretical guarantees, new insights, and performance improvements to widely-adopted Bayesian models.

References

- [1] P. Algikar and L. Mili. Robust gaussian process regression with huber likelihood, 2023. URL <https://arxiv.org/abs/2301.07858>.
- [2] M. Altamirano, F.-X. Briol, and J. Knoblauch. Robust and conjugate gaussian process regression. *arXiv preprint arXiv:2311.00463*, 2023.
- [3] S. Ament and C. Gomes. On the optimality of backward regression: Sparse recovery and subset selection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5599–5603, 2021. doi: 10.1109/ICASSP39728.2021.9415082.
- [4] S. Ament and C. Gomes. Generalized matching pursuits for the sparse optimization of separable objectives. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5553–5557, 2022. doi: 10.1109/ICASSP43922.2022.9747510.
- [5] S. Ament and M. O’Neil. Accurate and efficient numerical calculation of stable densities via optimized quadrature and asymptotics. *Statistics and Computing*, 28:171–185, 2018.
- [6] S. Ament, M. Amsler, D. R. Sutherland, M.-C. Chang, D. Guevarra, A. B. Connolly, J. M. Gregoire, M. O. Thompson, C. P. Gomes, and R. B. van Dover. Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams. *Science Advances*, 7(51): eabg4930, 2021. doi: 10.1126/sciadv.abg4930. URL <https://www.science.org/doi/abs/10.1126/sciadv.abg4930>.
- [7] S. Ament, S. Daulton, D. Eriksson, M. Balandat, and E. Bakshy. Unexpected improvements to expected improvement for bayesian optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 20577–20612. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/419f72cbd568ad62183f8132a3605a2a-Paper-Conference.pdf.
- [8] S. E. Ament and C. P. Gomes. Sparse bayesian learning via stepwise regression. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 264–274. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/ament21a.html>.
- [9] D. Andrade and A. Takeda. Robust Gaussian process regression with the trimmed marginal likelihood. In R. J. Evans and I. Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 67–76. PMLR, 31 Jul–04 Aug 2023. URL <https://proceedings.mlr.press/v216/andrade23a.html>.
- [10] P. Awasthi, A. Das, W. Kong, and R. Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=VHmdFFPy4U_u.
- [11] E. Bakshy, L. Dworkin, B. Karrer, K. Kashin, B. Letham, A. Murthy, and S. Singh. Ae: A domain-agnostic platform for adaptive experimentation. In *NeurIPS Workshop on Systems for Machine Learning*, 2018.
- [12] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020.
- [13] G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.
- [14] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus. Mathematique*, 346(9-10):589–592, 2008.
- [15] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data*, pages 2201–2206, 2016.

- [16] P. Congdon. Representing spatial dependence and spatial discontinuity in ecological epidemiology: a scale mixture approach. *Stochastic Environmental Research and Risk Assessment*, 31(2): 291–304, 2017.
- [17] A. Daemi, H. Kodamana, and B. Huang. Gaussian process modelling with gaussian mixture likelihood. *Journal of Process Control*, 81:209–220, 2019.
- [18] A. Das and D. Kempe. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19(3): 1–34, 2018.
- [19] L. C. W. Dixon. The global optimization problem: an introduction. *Towards Global Optimization 2*, pages 1–15, 1978.
- [20] K. Dong, D. Eriksson, H. Nickisch, D. Bindel, and A. G. Wilson. Scalable log determinants for gaussian process kernel learning. *Advances in Neural Information Processing Systems*, 30, 2017.
- [21] E. R. Elenberg, R. Khanna, A. G. Dimakis, and S. Negahban. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568, 2018.
- [22] P. I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.
- [23] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. GPyTorch: Blackbox Matrix-matrix Gaussian Process Inference with GPU Acceleration. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS’18*, pages 7587–7597, USA, 2018. Curran Associates Inc.
- [24] R. Garnett. *Bayesian Optimization*. Cambridge University Press, 2023. to appear.
- [25] J. Gerritsma, R. Onnink, and A. Versluis. Yacht hydrodynamics. UCI Machine Learning Repository, 1981. DOI: <https://doi.org/10.24432/C5XG7R>.
- [26] P. W. Goldberg, C. K. I. Williams, and C. Bishop. Regression with input-dependent noise: A gaussian process treatment. In *Advances in Neural Information Processing Systems*, volume 10, January 1998.
- [27] J. Harrison, David and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.
- [28] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [29] C. Hvarfner, E. Hellsten, F. Hutter, and L. Nardi. Self-correcting bayesian optimization through bayesian active learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [30] P. Jylänki, J. Vanhatalo, and A. Vehtari. Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12(11), 2011.
- [31] O. Karaca, D. Tihanyi, and M. Kamgarpour. Performance guarantees of forward and reverse greedy algorithms for minimizing nonsupermodular nonsubmodular functions on a matroid. *Operations Research Letters*, 49(6):855–861, 2021.
- [32] M. Kelly, R. Longjohn, and K. Nottingham. The uci machine learning repository. URL <https://archive.ics.uci.edu>, 2023.
- [33] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, pages 393–400, 2007.
- [34] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.
- [35] M. Kuss. *Gaussian process models for robust regression, classification, and reinforcement learning*. PhD thesis, Technische Universität Darmstadt, 2006.
- [36] Z.-Z. Li, L. Li, and Z. Shao. Robust gaussian process regression based on iterative trimming. *Astronomy and Computing*, 36:100483, 2021. ISSN 2213-1337. doi: <https://doi.org/10.1016/j.ascom.2021.100483>. URL <https://www.sciencedirect.com/science/article/pii/S2213133721000378>.
- [37] J. A. Lin, S. Ament, M. Balandat, and E. Bakshy. Scaling gaussian processes for learning curve prediction via latent kronecker structure, 2024. URL <https://arxiv.org/abs/2410.09239>.

- [38] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(1–3):503–528, Aug. 1989. ISSN 0025-5610.
- [39] F. Locatello, R. Khanna, M. Tschannen, and M. Jaggi. A Unified Optimization View on Generalized Matching Pursuit and Frank-Wolfe. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 860–868. PMLR, 20–22 Apr 2017. URL <https://proceedings.mlr.press/v54/locatello17a.html>.
- [40] S. Lotfi, P. Izmailov, G. Benton, M. Goldblum, and A. G. Wilson. Bayesian model selection, the marginal likelihood, and generalization, 2023. URL <https://arxiv.org/abs/2202.11678>.
- [41] R. Martinez-Cantin, M. McCourt, and K. Tee. Robust bayesian optimization with student-t likelihood, 2017.
- [42] N. Maus, K. Wu, D. Eriksson, and J. Gardner. Discovering many diverse solutions with Bayesian optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 206, pages 1779–1798. PMLR, Apr. 2023.
- [43] A. Naish-Guzman and S. Holden. Robust regression with twinned gaussian processes. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/2ab56412b1163ee131e1246da0955bd1-Paper.pdf.
- [44] R. M. Neal. Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Technical Report 9702, Department of Statistics, University of Toronto, 1997.
- [45] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.
- [46] R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3): 291–297, 1997.
- [47] C. Park, D. J. Borth, N. S. Wilson, C. N. Hunter, and F. J. Friedersdorf. Robust gaussian process regression with a bias model. *Pattern Recognition*, 124:108444, 2022.
- [48] R. Ranjan, B. Huang, and A. Fatehi. Robust gaussian process modeling using em algorithm. *Journal of Process Control*, 42:125–136, 2016.
- [49] C. E. Rasmussen, C. K. Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.
- [50] C. Riis, F. Antunes, F. Hüttel, C. Lima Azevedo, and F. Pereira. Bayesian active learning with fully bayesian gaussian processes. *Advances in Neural Information Processing Systems*, 35: 12141–12153, 2022.
- [51] A. Shah, A. Wilson, and Z. Ghahramani. Student-t processes as alternatives to gaussian processes. In *Artificial intelligence and statistics*, pages 877–885. PMLR, 2014.
- [52] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, page 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.
- [53] S. Sundararajan and S. Keerthi. Predictive approaches for choosing hyperparameters in gaussian processes. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/e8fd4a8a5bab2b3785d794ab51fef55c-Paper.pdf.
- [54] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.
- [55] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.
- [56] M. E. Tipping and N. D. Lawrence. Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(1):123–141, 2005.
- [57] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

- [58] A. Tsanas and A. Xifara. Energy Efficiency. UCI Machine Learning Repository, 2012. DOI: <https://doi.org/10.24432/C51307>.
- [59] J. Vanhatalo, P. Jylänki, and A. Vehtari. Gaussian process regression with student-t likelihood. In *Advances in Neural Information Processing Systems*, volume 22, 2009.
- [60] Z. Wang, C. Gehring, P. Kohli, and S. Jegelka. Batched large-scale bayesian optimization in high-dimensional spaces. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 84 of *PMLR*, pages 745–754. JMLR, Mar. 2018.
- [61] L. Wasserman. Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107, 2000.
- [62] D. Wipf and S. Nagarajan. A new view of automatic relevance determination. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/9c01802ddb981e6bcfbec0f0516b8e35-Paper.pdf.
- [63] I.-C. Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 1998. DOI: <https://doi.org/10.24432/C5PK67>.

A Additional Details on the Model

Algorithm 2 below is the “backward” variant of Algorithm 1 from Sec 4. As its name suggests, the main difference compared to the “forward” variant is that rather than building up a set of “outliers”, it starts from a (typically large) set of “outliers” and iteratively *removes* those data points from the set that have the smallest inferred data-point-dependent noise variance ρ_i .

While we have not derived theoretical guarantees for this “backward” version, we have found it to generally behave similarly to the “forward” version in terms of performance and robustness. One empirical observation from our studies is that while the “forward” version tends to perform slightly better than the “backward” version if there are only few outliers, the opposite is true if the outlier frequency is very high. This behavior is rather intuitive and illustrates that relevance pursuit is particularly well-suited to identify sparse, low-cardinality subsets (note that in the “backward” variant under large corruptions, the uncorrupted data points can be viewed as the sparse subset that needs to be identified).

Algorithm 2 Relevance Pursuit (Backward Algorithm)

Require: \mathbf{X}, \mathbf{y} , schedule $\mathcal{K} = (k_1, k_2, \dots)$

Initialize $\mathcal{S}_0^c \subseteq \{1, \dots, n\}$ (typically $\mathcal{S}_0^c = \{1, \dots, n\}$)

for k_i in \mathcal{K} **do**

Optimize ML: $\rho_{\mathcal{S}_i^c}^* \leftarrow \arg \max_{\rho_{\mathcal{S}_i^c}} \mathcal{L}(\rho_{\mathcal{S}_i^c})$, where $\rho_{\mathcal{S}_i^c} = \{\rho : \rho_j = 0, \forall j \notin \mathcal{S}_i^c\}$

Compute the set \mathcal{R}_i containing the k_i elements of \mathcal{S}_i^c with smallest inferred variance:

$\mathcal{R}_i \leftarrow \{j_i^1, \dots, j_i^{k_i}\}$ where $j_i^l \in \mathcal{S}_i^c$ such that $\rho_{\mathcal{S}_i^c}^*(j_i^l) \leq \rho_{\mathcal{S}_i^c}^*(j_i^{l'})$ for $l < l'$

$\mathcal{S}_{i+1}^c \leftarrow \mathcal{S}_i^c \setminus \mathcal{R}_i$

$\mathcal{S}_i \leftarrow \{1, \dots, n\} \setminus \mathcal{S}_i^c$ for each k_i

Compute the marginal likelihood $p(\mathcal{S}_i | \mathbf{X}, \mathbf{y}) \approx p(\mathcal{S}_i, \rho_{\mathcal{S}_i}^* | \mathbf{X}, \mathbf{y})$

$\mathcal{S}^* \leftarrow \arg \max_{\mathcal{S}_i} p(\mathcal{S}_i | \mathbf{X}, \mathbf{y}) p(\mathcal{S}_i)$.

Return $\mathcal{S}^*, \rho_{\mathcal{S}^*}^*$.

B Additional Background on the Theory

B.1 Submodular Functions

Das and Kempe [18] showed that the subset selection problem of regression features with an R^2 objective satisfies a weak submodularity property, which can be invoked to prove approximation guarantees for the greedy maximization of the objective. Elenberg et al. [21] generalized this work by proving that any log likelihood function exhibiting restricted strong concavity gives rise to the weak submodularity of the associated subset selection problem, which can be invoked to prove approximation guarantees for the greedy algorithm. Karaca et al. [31] contains a guarantee for the backward algorithm applied to the maximization of submodular set functions.

Submodularity has also found application to Gaussian process models. For a sensor placement problem, Krause et al. [34] proved that the mutual information (MI) criterion, capturing the reduction in uncertainty in the entire search space, can be a submodular function. In this case, MI is not monotonic everywhere, but monotonic for small sets ($2k$) of sensors, which is sufficient to apply Nemhauser’s guarantee for sparse sets of sensors up to size k [45]. Srinivas et al. [52] used the submodularity of the joint entropy in order to prove regret bounds for the convergence of a GP-based BO algorithm using the upper-confidence bound acquisition function.

Elenberg et al. [21] proved that any log likelihood function exhibiting restricted strong concavity and smoothness implies the weak submodularity of the associated subset selection problem.

Definition 9 (Submodularity Ratios [21]). *Let $\mathcal{A}, \mathcal{B} \subset [n]$ be two disjoint sets, and $f : 2^{[n]} \rightarrow \mathbb{R}$. The submodularity ratio of \mathcal{B} with respect to \mathcal{A} is defined by*

$$\gamma_{\mathcal{B}, \mathcal{A}} = \sum_{i \in \mathcal{A}} (f(\mathcal{B} \cup \{i\}) - f(\mathcal{B})) / (f(\mathcal{B} \cup \mathcal{A}) - f(\mathcal{B}))$$

The submodularity ratio of a set \mathcal{C} with respect to an integer k is defined by

$$\gamma_{\mathcal{C},k} = \min_{\mathcal{B},\mathcal{A}} \gamma_{\mathcal{B},\mathcal{A}} \quad \text{such that} \quad \mathcal{A} \cap \mathcal{B} = \emptyset, \quad \mathcal{B} \subseteq \mathcal{C}, \quad \text{and} \quad |\mathcal{A}| \leq k.$$

Then given $\gamma > 0$, a function is γ -weakly submodular at a set \mathcal{C} with respect to k if $\gamma_{\mathcal{C},k} \geq \gamma$.

Theorem 10 (Weak Submodularity via RSC [21]). *The submodularity ratio $\gamma_{\mathcal{S},k}$ can be bound below using the restricted convexity and smoothness parameters $m_{|\mathcal{S}|+k}$ and $M_{|\mathcal{S}|+k}$,*

$$\gamma_{\mathcal{S},k} \geq m_{|\mathcal{S}|+k} / M_{|\mathcal{S}|+k}.$$

Theorem 11 (OMP Approximation Guarantee [21]). *Let $\mathbf{x}_{\text{OMP}}(r)$ be the r -sparse vector selected by OMP, and $\mathbf{x}_{\text{OPT}}(r) = \arg \min_{\|\mathbf{x}\|_0=r} f(\mathbf{x})$ be the optimal r -sparse vector. Then*

$$f(\mathbf{x}_{\text{OMP}}(r)) \geq \left(1 - e^{-m_{2r}/M_{2r}}\right) f(\mathbf{x}_{\text{OPT}}(r)),$$

where m_{2r}, M_{2r} are the restricted strong convexity and smoothness parameters of f , respectively.

C Theoretical Results and Proofs

Lemma 1. [Optimal Robust Variances] *Let $\mathcal{D}_{\setminus i} = \{(\mathbf{x}_j, y_j) : j \neq i\}$, $\boldsymbol{\rho} = \boldsymbol{\rho}_{\setminus i} + \rho_i \mathbf{e}_i$, where $\boldsymbol{\rho}, \boldsymbol{\rho}_{\setminus i} \in \mathbb{R}_+^n$, $[\boldsymbol{\rho}_{\setminus i}]_i = 0$, and \mathbf{e}_i is the i th canonical basis vector. Then keeping $\boldsymbol{\rho}_{\setminus i}$ fixed,*

$$\rho_i^* = \arg \max_{\rho_i} \mathcal{L}(\boldsymbol{\rho}_{\setminus i} + \rho_i \mathbf{e}_i) = \left[(y_i - \mathbb{E}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}])^2 - \mathbb{V}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}] \right]_+, \quad (3)$$

where $y(\mathbf{x}_i) = f(\mathbf{x}_i) + \epsilon_i$. These quantities can be expressed as functions of $\boldsymbol{\Sigma}^{-1} = (\mathbf{K} + \mathbf{D}_{\sigma^2 + \boldsymbol{\rho}})^{-1}$:

$$\mathbb{E}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}]^2 = y_i - [\boldsymbol{\Sigma}^{-1} \mathbf{y}]_i / [\boldsymbol{\Sigma}^{-1}]_{ii}, \quad \text{and} \quad \mathbb{V}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}] = 1 / [\boldsymbol{\Sigma}^{-1}]_{ii},$$

where $\mathbf{D}_{\sigma^2 + \boldsymbol{\rho}}$ is a diagonal matrix whose entries are $\sigma^2 + \rho$.

Proof. First, we partition the covariance matrix $\mathbf{K} + \sigma^2 \mathbf{I} + \mathbf{D}_{\boldsymbol{\rho}}$ to separate the effect of ρ_i and use the block matrix inverse

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{K} + \mathbf{D}_{\boldsymbol{\rho} + \sigma^2})^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{\setminus i}^{-1} + \mathbf{u} \beta_i \mathbf{u}^\top & -\mathbf{u} \beta_i \\ -\mathbf{u}^\top \beta_i & \beta_i \end{bmatrix}, \quad (5)$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{\setminus i} &= k(\mathbf{X}_{\setminus i}, \mathbf{X}_{\setminus i}) + \mathbf{D}_{\boldsymbol{\rho}_{\setminus i} + \sigma^2}, \\ \mathbf{u} &= \boldsymbol{\Sigma}_{\setminus i}^{-1} k(\mathbf{X}_{\setminus i}, \mathbf{x}_i), \quad \text{and} \\ \beta_i &= \left([k(\mathbf{x}_i, \mathbf{x}_i) + \sigma^2 + \rho_i] - k(\mathbf{x}_i, \mathbf{X}_{\setminus i}) \boldsymbol{\Sigma}_{\setminus i}^{-1} k(\mathbf{X}_{\setminus i}, \mathbf{x}_i) \right)^{-1}. \end{aligned} \quad (6)$$

Quadratic Term With the expression for the inverse of $\boldsymbol{\Sigma}$ above, we can write the quadratic term of the log likelihood as

$$\begin{aligned} \mathbf{y}^\top (\mathbf{K} + \mathbf{D}_{\boldsymbol{\rho} + \sigma^2})^{-1} \mathbf{y} &= \mathbf{y}_{\setminus i}^\top (\boldsymbol{\Sigma}_{\setminus i}^{-1} + \mathbf{u} \beta_i \mathbf{u}^\top) \mathbf{y}_{\setminus i} - 2 \mathbf{y}_{\setminus i}^\top \mathbf{u} \beta_i y_i + y_i^2 \beta_i \\ &= \mathbf{y}_{\setminus i}^\top \boldsymbol{\Sigma}_{\setminus i}^{-1} \mathbf{y}_{\setminus i} + \beta_i (\mathbf{y}_{\setminus i}^\top \mathbf{u} - y_i)^2. \end{aligned} \quad (7)$$

Determinant Term The determinant of a block matrix is given by

$$\left| \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \right| = |\mathbf{A}| |\mathbf{D} - \mathbf{C} \mathbf{A}^{-1} \mathbf{B}|. \quad (8)$$

Applying this identity to $\boldsymbol{\Sigma} = (\mathbf{K} + \mathbf{D}_{\boldsymbol{\rho} + \sigma^2})$, we get

$$|\mathbf{K} + \mathbf{D}_{\boldsymbol{\rho} + \sigma^2}| = |\boldsymbol{\Sigma}_{\setminus i}| \beta_i^{-1}. \quad (9)$$

Log Marginal Likelihood

$$2(\mathcal{L}(\boldsymbol{\rho}) - \mathcal{L}(\boldsymbol{\rho}_{\setminus i})) = -\beta_i(\mathbf{y}_{\setminus i}^\top \mathbf{u} - y_i)^2 - \log(\beta_i^{-1}). \quad (10)$$

Noting that $\partial_{\rho_i} \beta_i = -\beta_i^2$, the derivative of the difference in log marginal likelihood w.r.t. ρ_i , is

$$\partial_{\rho_i} 2(\mathcal{L}(\boldsymbol{\rho}) - \mathcal{L}(\boldsymbol{\rho}_{\setminus i})) = (\mathbf{y}_{\setminus i}^\top \mathbf{u} - y_i)^2 \beta_i^2 - \beta_i. \quad (11)$$

While $\beta_i = 0$ is a root of the derivative, we ignore this solution since β_i is never zero when $\sigma^2 > 0$. Therefore, the remaining stationary point is $\beta_i^{-1} = (\mathbf{y}_{\setminus i}^\top \mathbf{u} - y_i)^2$. Since we constrain $\rho \geq 0$, this point might not always be attainable. However, because there is only a single stationary point with respect to β_i when $\sigma^2 > 0$, and β_i is a strictly decreasing function of ρ_i , it follows that the marginal likelihood is monotonic as a function of ρ_i to both the left and the right of the stationary point. Therefore, the optimal constraint ρ_i is simply the optimal unconstrained value, projected into the feasible space. In particular, solving $\beta_i^{-1} = (\mathbf{y}_{\setminus i}^\top \mathbf{u} - y_i)^2$ for ρ_i and projecting to the non-negative half-line, we get

$$\rho_i = \left[\left(\mathbf{y}_{\setminus i}^\top \mathbf{u} - y_i \right)^2 - \left(k(\mathbf{x}_i, \mathbf{x}_i) + \sigma^2 - k(\mathbf{x}_i, \mathbf{X}_{\setminus i}) \boldsymbol{\Sigma}_{\setminus i}^{-1} k(\mathbf{X}_{\setminus i}, \mathbf{x}_i) \right) \right]_+. \quad (12)$$

Lastly, note

$$\begin{aligned} \mathbf{y}_{\setminus i}^\top \mathbf{u} - y_i &= \mathbf{y}_{\setminus i}^\top (k(\mathbf{X}_{\setminus i}, \mathbf{X}_{\setminus i}) + \mathbf{D}_{\sigma^2 + \rho_{\setminus i}})^{-1} k(\mathbf{X}_{\setminus i}, \mathbf{x}_i) - y_i = \mathbb{E}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}] - y_i, \\ k(\mathbf{x}_i, \mathbf{x}_i) - k(\mathbf{x}_i, \mathbf{X}_{\setminus i}) \boldsymbol{\Sigma}_{\setminus i}^{-1} k(\mathbf{X}_{\setminus i}, \mathbf{x}_i) + \sigma^2 &= \mathbb{V}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}]. \end{aligned} \quad (13)$$

As stated by Rasmussen et al. [49] (P. 117, Eq. 5.12), originally shown by Sundararajan and Keerthi [53], these quantities can be expressed as simple functions of $\boldsymbol{\Sigma}^{-1}$:

$$\begin{aligned} \mathbb{E}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}]^2 &= y_i - [\boldsymbol{\Sigma}^{-1} \mathbf{y}]_i / [\boldsymbol{\Sigma}^{-1}]_{ii} \\ \mathbb{V}[y(\mathbf{x}_i) | \mathcal{D}_{\setminus i}] &= 1 / [\boldsymbol{\Sigma}^{-1}]_{ii}. \end{aligned} \quad (14)$$

Therefore, all LOO predictive values can be computed in $\mathcal{O}(n^3)$ or faster, if an inducing point method is used for \mathbf{K} . \square

The following is a preliminary result for our analysis of the log marginal likelihood w.r.t. $\boldsymbol{\rho}$.

Lemma 12. *The gradient and Hessian of the log marginal likelihood \mathcal{L} with respect to $\boldsymbol{\rho}$ are given by*

$$-2\nabla_{\boldsymbol{\rho}}[\mathcal{L}] = \text{diag}(\mathbf{K}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top), \quad \text{and} \quad -2\mathbf{H}_{\boldsymbol{\rho}}[\mathcal{L}] = (2\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}^{-1}) \odot \mathbf{K}^{-1},$$

where $\mathbf{K} = \mathbf{K}_0 + \mathbf{D}_{\boldsymbol{\rho}}$ for some base covariance matrix \mathbf{K}_0 and $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$.

Proof. Let $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$. Regarding the gradient, note that $\partial_{\rho_i} \mathbf{K} = \mathbf{e}_i \mathbf{e}_i^\top$, where \mathbf{e}_i is the canonical basis vector with a one as the i th element, and based on Equation 5.9 of Rasmussen et al. [49],

$$\begin{aligned} -2\partial_{\rho_i}[\mathcal{L}] &= \text{tr}((\mathbf{K}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top) \partial_{\rho_i} \mathbf{K}) \\ &= \text{tr}((\mathbf{K}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top) \mathbf{e}_i \mathbf{e}_i^\top) \\ &= \mathbf{e}_i^\top (\mathbf{K}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top) \mathbf{e}_i \\ &= (\mathbf{K}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top)_{ii}. \end{aligned}$$

Regarding the second derivatives, according to Dong et al. [20],

$$\begin{aligned} \partial_{\theta_i} \partial_{\theta_j} [\log |\mathbf{K}|] &= \text{tr}(\mathbf{K}^{-1} [\partial_{\theta_i} \partial_{\theta_j} \mathbf{K}] - \mathbf{K}^{-1} [\partial_{\theta_i} \mathbf{K}] \mathbf{K}^{-1} [\partial_{\theta_j} \mathbf{K}]) \\ \partial_{\theta_i} \partial_{\theta_j} [\mathbf{y}^\top \mathbf{K}^{-1} \mathbf{y}] &= 2\boldsymbol{\alpha}^\top [\partial_{\theta_i} \mathbf{K}] \mathbf{K}^{-1} [\partial_{\theta_j} \mathbf{K}] \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top [\partial_{\theta_i} \partial_{\theta_j} \mathbf{K}] \boldsymbol{\alpha}. \end{aligned}$$

Therefore,

$$\begin{aligned} -2\partial_{\rho_i} \partial_{\rho_j} \mathcal{L} &= \text{tr}((\mathbf{K}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top) [\partial_{\rho_i} \partial_{\rho_j} \mathbf{K}] - (\mathbf{K}^{-1} - 2\boldsymbol{\alpha}\boldsymbol{\alpha}^\top) ([\partial_{\rho_i} \mathbf{K}] \mathbf{K}^{-1} [\partial_{\rho_j} \mathbf{K}])) \\ &= \text{tr}((2\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}^{-1}) ([\mathbf{K}^{-1}]_{ij} \mathbf{e}_i \mathbf{e}_j^\top)) \\ &= \text{tr}(\mathbf{e}_j^\top (2\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}^{-1}) \mathbf{e}_i) [\mathbf{K}^{-1}]_{ij} \\ &= [2\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}^{-1}]_{ij} [\mathbf{K}^{-1}]_{ij} \\ &= [(2\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}^{-1}) \odot \mathbf{K}^{-1}]_{ij}, \end{aligned}$$

since $[\partial_{\rho_i} \partial_{\rho_j} \mathbf{K}] = \mathbf{0}$. The third equality is due to the invariance of the trace to circular shifts of its argument. The fourth equality is due to the symmetry of the matrix in brackets. \square

C.1 Strong Convexity and Smoothness of the Reparameterized Robust Marginal Likelihood

Here, we re-parameterize $\boldsymbol{\rho}(\mathbf{s}) = \text{diag}(\mathbf{K}_0) \odot ((1 - \mathbf{s})^{-1} - 1)$, and attain strong convexity for all inputs \mathbf{s} , if conditions on the eigenvalues of the covariance matrix and the norm of the data vector $\|\mathbf{y}\|$ are met. The convexity result is surprising in two ways: the negative log marginal likelihood of GPs is generally a non-convex function, and in addition, the negative log likelihoods of many alternative robust regression methods like the Student- t likelihood or α -stable likelihoods are non-convex, and even Huber's proposal is non-*strongly*-convex.

Lemma 4. [Reparameterized Hessian] Let $\mathbf{K}_\mathbf{s} = k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I} + \mathbf{D}_{\boldsymbol{\rho}(\mathbf{s})}$, $\hat{\mathbf{K}}_\mathbf{s} = \text{diag}(\mathbf{K}_\mathbf{s})^{-1/2} \mathbf{K}_\mathbf{s} \text{diag}(\mathbf{K}_\mathbf{s})^{-1/2}$, and $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{K}}_\mathbf{s}^{-1} \text{diag}(\mathbf{K}_\mathbf{s})^{-1/2} \mathbf{y}$. Then

$$\mathbf{H}_\mathbf{s}[-2\mathcal{L}(\boldsymbol{\rho}(\mathbf{s}))] = \mathbf{D}_{1-\mathbf{s}}^{-1} \left[2 \left(\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) + 2 \text{diag}(\hat{\mathbf{K}}^{-1}) - (\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}) \right] \mathbf{D}_{1-\mathbf{s}}^{-1}.$$

Proof. Using the chain-rule, the Hessian $\mathbf{H}_\mathbf{s}[\mathcal{L}]$ can be expressed as a function of the Jacobian $\mathbf{J}_\mathbf{s}[\boldsymbol{\rho}(\mathbf{s})] = \mathbf{D}_{\partial \boldsymbol{\rho}(\mathbf{s})}$, which is diagonal since $\boldsymbol{\rho}(\mathbf{s})$ is an element-wise function, and the second derivatives $\partial_s^2 \rho(s_i)$. Then

$$\begin{aligned} -2\mathbf{H}_\mathbf{s}[\mathcal{L}] &= -\mathbf{J}_\mathbf{s}[\boldsymbol{\rho}]^\top \mathbf{H}_\rho[2 \log \mathcal{L}] \mathbf{J}_\mathbf{s}[\boldsymbol{\rho}] + \mathbf{D}_{\nabla_\rho[-2\mathcal{L}]} \mathbf{D}_{\partial_s^2 \rho} \\ &= \mathbf{D}_{\boldsymbol{\rho}'(\mathbf{s})} [(2\boldsymbol{\alpha} \boldsymbol{\alpha}^\top - \mathbf{K}^{-1}) \odot \mathbf{K}^{-1}] \mathbf{D}_{\boldsymbol{\rho}'(\mathbf{s})} + \text{diag}(\mathbf{K}^{-1}) \mathbf{D}_{\boldsymbol{\rho}''} - \mathbf{D}_{\boldsymbol{\rho}'(\mathbf{s})} (\mathbf{K}^{-1} \odot \mathbf{K}^{-1}) \mathbf{D}_{\boldsymbol{\rho}'}, \end{aligned}$$

where we substituted the relevant expressions from Lemma 12. Further substituting $\boldsymbol{\rho}(\mathbf{s})_i = [\mathbf{K}_0]_{ii}[(1 - s_i)^{-1} - 1]$, $\boldsymbol{\rho}'(\mathbf{s})_i = [\mathbf{K}_0]_{ii}(1 - s_i)^{-2}$, and $\boldsymbol{\rho}''(\mathbf{s})_i = 2[\mathbf{K}_0]_{ii}(1 - s_i)^{-3}$, noting that $\mathbf{K} = \mathbf{K}_0 + \text{diag}(\mathbf{K}_0)[(1 - \mathbf{s})^{-1} - 1] = (\mathbf{K}_0 - \text{diag}(\mathbf{K}_0)) + \text{diag}(\mathbf{K}_0)(1 - \mathbf{s})^{-1}$, and algebraic manipulation finish the proof. \square

Lemma 5. [Strong Convexity via Eigenvalue Condition] Let $\hat{\mathbf{K}}_\mathbf{s}$ as in Lemma 4. Then $\mathbf{H}_\mathbf{s} \succ m \mathbf{I}$ if

$$\lambda_{\min} \hat{\lambda}_{\min}^2 \frac{(2\hat{\lambda}_{\max}^{-1} - \hat{\lambda}_{\min}^{-2} - m)}{2(1 - \lambda_{\min}/\lambda_{\max})} > \|\mathbf{y}\|_2^2, \quad (4)$$

where $\lambda_{\min, \max}$ (resp. $\hat{\lambda}_{\min, \max}$) are the smallest and largest eigenvalues of $\mathbf{K}_\mathbf{s}$, respectively $\hat{\mathbf{K}}_\mathbf{s}$.

Proof. We seek to lower-bound the smallest eigenvalue of the Hessian matrix, which—for twice-continuously-differentiable problems—is equivalent to lower and upper bounds of the problem's Hessian matrix. Starting with the result of Lemma 4,

$$\begin{aligned} \mathbf{H}_\mathbf{s}[-2\mathcal{L}(\boldsymbol{\rho}(\mathbf{s}))] &= \mathbf{D}_{1-\mathbf{s}}^{-1} \left[2 \left(\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) + 2 \text{diag}(\hat{\mathbf{K}}^{-1}) - (\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}) \right] \mathbf{D}_{1-\mathbf{s}}^{-1} \\ &\succeq 2 \left(\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) + 2 \text{diag}(\hat{\mathbf{K}}^{-1}) - (\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}). \end{aligned}$$

Now, we bound each of the three additive terms independently from below.

Term 1:

$$2 \text{diag}(\hat{\mathbf{K}}^{-1}) \succeq 2\lambda_{\min}(\hat{\mathbf{K}}^{-1})\mathbf{I} = 2\lambda_{\max}(\hat{\mathbf{K}})^{-1}\mathbf{I}.$$

The first inequality comes from \mathbf{K} being positive definite, and the absolute value of the diagonal of a matrix, which is already positive for positive definite matrices, being lower bounded by the minimum eigenvalue of the matrix. The last steps is a basic consequence of the eigenvalues of inverses matrices. Note that the eigenvalues of $\hat{\mathbf{K}}$ can be further bound by the eigenvalues of the original matrix \mathbf{K} :

$$\begin{aligned} \lambda_{\min}(\hat{\mathbf{K}}) &= \lambda_{\min}(\text{diag}(\mathbf{K})^{-1/2} \mathbf{K} \text{diag}(\mathbf{K})^{-1/2}) \\ &\geq \lambda_{\min}(\text{diag}(\mathbf{K})^{-1/2})^2 \lambda_{\min}(\mathbf{K}) \\ &= \lambda_{\min}(\text{diag}(\mathbf{K})^{-1}) \lambda_{\min}(\mathbf{K}) \\ &\geq \lambda_{\min}(\mathbf{K}^{-1}) \lambda_{\min}(\mathbf{K}) \\ &= \lambda_{\min}(\mathbf{K}) / \lambda_{\max}(\mathbf{K}). \end{aligned}$$

In a similar way, we can show that $\lambda_{\max}(\hat{\mathbf{K}}) \leq \lambda_{\max}(\mathbf{K}) / \lambda_{\min}(\mathbf{K})$, which implies $\lambda_{\max}(\hat{\mathbf{K}})^{-1} \geq \lambda_{\min}(\mathbf{K}) / \lambda_{\max}(\mathbf{K})$.

Term 2: Next, we have

$$-(\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}) \succeq -\lambda_{\max}(\hat{\mathbf{K}}^{-1})^2 \mathbf{I} = -\lambda_{\min}(\hat{\mathbf{K}})^{-2} \mathbf{I},$$

which is due to the Hadamard product being a sub-matrix of the Kronecker product of the same matrices, the largest eigenvalue of the former are bounded by the largest eigenvalue of the latter, which is the product of the largest eigenvalues of the constituent matrices.

Term 3: Lastly, note that

$$\begin{aligned} 2 \left(\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) &= 2 \mathbf{D}_{\hat{\boldsymbol{\alpha}}} (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \mathbf{D}_{\hat{\boldsymbol{\alpha}}} \\ &\succeq 2 \mathbf{D}_{\hat{\boldsymbol{\alpha}}} (\lambda_{\min}(\hat{\mathbf{K}}^{-1}) \mathbf{I} - \mathbf{I}) \mathbf{D}_{\hat{\boldsymbol{\alpha}}} \\ &= 2 (\lambda_{\min}(\hat{\mathbf{K}}^{-1}) - 1) \mathbf{D}_{\hat{\boldsymbol{\alpha}}}^2 \\ &= 2 (\lambda_{\max}(\hat{\mathbf{K}})^{-1} - 1) \mathbf{D}_{\hat{\boldsymbol{\alpha}}}^2 \\ &\succeq 2 (\lambda_{\min}(\mathbf{K}) / \lambda_{\max}(\mathbf{K}) - 1) \mathbf{D}_{\hat{\boldsymbol{\alpha}}}^2 \\ &\succeq 2 (\lambda_{\min}(\mathbf{K}) / \lambda_{\max}(\mathbf{K}) - 1) \|\hat{\boldsymbol{\alpha}}\|_\infty^2 \mathbf{I}, \end{aligned}$$

where the last inequality comes from the second to last lower bound of Term 3 being non-positive, and therefore, being able to lower bound it with the largest magnitude entry of $\hat{\boldsymbol{\alpha}}$.

Term 1 + 2 + 3: Putting together the inequalities for all terms, we get

$$\lambda_{\min}(\mathbf{H}_s) \geq 2\lambda_{\max}(\hat{\mathbf{K}})^{-1} - \lambda_{\min}(\hat{\mathbf{K}})^{-2} + 2(\lambda_{\min}(\mathbf{K}) / \lambda_{\max}(\mathbf{K}) - 1) \|\hat{\boldsymbol{\alpha}}\|_\infty^2,$$

where in slight abuse of notation, we let $\mathbf{H}_s = \mathbf{H}_s[-2\mathcal{L}]$ be the Hessian of the negative log likelihood.

Using the bound $\|\hat{\boldsymbol{\alpha}}\|_\infty \leq \|\hat{\boldsymbol{\alpha}}\|_2 \leq \lambda_{\min}(\hat{\mathbf{K}})^{-1} \|\hat{\mathbf{y}}\|$, where $\hat{\mathbf{y}} = \text{diag}(\mathbf{K})^{-1/2} \mathbf{y}$. Therefore, $\|\hat{\boldsymbol{\alpha}}\|_\infty \leq \lambda_{\min}(\hat{\mathbf{K}})^{-1} \lambda_{\min}(\mathbf{K})^{-1/2} \|\mathbf{y}\|_2 \leq \lambda_{\max}(\mathbf{K}) \lambda_{\min}(\mathbf{K})^{-3/2} \|\mathbf{y}\|_2$.

Finally, lower bounding the current lower bound by $m > 0$ yields a sufficient condition for the convexity at \mathbf{s} .

$$2\lambda_{\max}(\hat{\mathbf{K}})^{-1} - \lambda_{\min}(\hat{\mathbf{K}})^{-2} + 2(\lambda_{\min}(\mathbf{K}) / \lambda_{\max}(\mathbf{K}) - 1) \lambda_{\min}(\hat{\mathbf{K}})^{-2} \lambda_{\min}(\mathbf{K})^{-1} \|\mathbf{y}\|_2^2 > m.$$

Re-arranging, we attain

$$\lambda_{\min} \hat{\lambda}_{\min}^2 \frac{(2\hat{\lambda}_{\max}^{-1} - \hat{\lambda}_{\min}^{-2} - m)}{2(1 - \lambda_{\min} / \lambda_{\max})} > \|\mathbf{y}\|_2^2, \quad (15)$$

which is a non-trivial guarantee when $2\hat{\lambda}_{\max}^{-1} - \hat{\lambda}_{\min}^{-2} - m > 0$. \square

Lemma 13. [Smoothness via Eigenvalue Condition] Let $\hat{\mathbf{K}}$ as in Lemma 4. Suppose that $\|\mathbf{s}\|_\infty \leq s_{\max}$ and

$$\lambda_{\min} \hat{\lambda}_{\min}^2 \frac{(M(1 - s_{\max})^2 + \hat{\lambda}_{\min}^{-2} - 2\hat{\lambda}_{\max}^{-1})}{2(\lambda_{\max} / \lambda_{\min} - 1)} > \|\mathbf{y}\|_2^2, \quad (16)$$

where $\lambda_{\min, \max}$ (resp. $\hat{\lambda}_{\min, \max}$) are the smallest and largest eigenvalues of \mathbf{K} (resp. $\hat{\mathbf{K}}$) respectively. Then $\mathbf{H}_s \prec M$. This is a non-trivial guarantee when $M(1 - s_{\max})^2 + \hat{\lambda}_{\min}^{-2} - 2\hat{\lambda}_{\max}^{-1} > 0$.

Proof. We now derive an equivalent upper bound for the largest eigenvalue of the Hessian. Starting with the result of Lemma 4,

$$\begin{aligned} \mathbf{H}_s[-2\mathcal{L}(\boldsymbol{\rho}(\mathbf{s}))] &= \mathbf{D}_{1-\mathbf{s}}^{-1} \left[2 \left(\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) + 2 \text{diag}(\hat{\mathbf{K}}^{-1}) - (\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}) \right] \mathbf{D}_{1-\mathbf{s}}^{-1} \\ &\preceq \|(1 - \mathbf{s})^{-1}\|_\infty^2 \left[2 \left(\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) + 2 \text{diag}(\hat{\mathbf{K}}^{-1}) - (\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}) \right]. \end{aligned}$$

Therefore, it becomes immediately apparent that we will need to introduce an upper bound on $\|(1 - \mathbf{s})^{-1}\|_\infty^2$, which is a restriction on the domain that $\boldsymbol{\rho}$ can take. Proceeding in a similar way as above, we bound the three terms in square brackets, now from above.

Term 1:

$$2 \operatorname{diag}(\hat{\mathbf{K}}^{-1}) \preceq 2\lambda_{\max}(\hat{\mathbf{K}}^{-1})\mathbf{I} = 2\lambda_{\min}(\hat{\mathbf{K}})^{-1}\mathbf{I}.$$

Term 2:

$$-(\hat{\mathbf{K}}^{-1} \odot \hat{\mathbf{K}}^{-1}) \preceq -\lambda_{\min}(\hat{\mathbf{K}}^{-1})^2\mathbf{I} = -\lambda_{\max}(\hat{\mathbf{K}})^{-2}\mathbf{I}.$$

Term 3:

$$\begin{aligned} 2 \left(\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^\top \odot (\hat{\mathbf{K}}^{-1} - \mathbf{I}) \right) &= 2\mathbf{D}_{\hat{\boldsymbol{\alpha}}}(\hat{\mathbf{K}}^{-1} - \mathbf{I})\mathbf{D}_{\hat{\boldsymbol{\alpha}}} \\ &\preceq 2\mathbf{D}_{\hat{\boldsymbol{\alpha}}}(\lambda_{\max}(\hat{\mathbf{K}}^{-1})\mathbf{I} - \mathbf{I})\mathbf{D}_{\hat{\boldsymbol{\alpha}}} \\ &= 2(\lambda_{\max}(\hat{\mathbf{K}}^{-1}) - 1)\mathbf{D}_{\hat{\boldsymbol{\alpha}}}^2 \\ &= 2(\lambda_{\min}(\hat{\mathbf{K}})^{-1} - 1)\mathbf{D}_{\hat{\boldsymbol{\alpha}}}^2 \\ &\preceq 2(\lambda_{\max}(\mathbf{K})/\lambda_{\min}(\mathbf{K}) - 1)\mathbf{D}_{\hat{\boldsymbol{\alpha}}}^2 \\ &\preceq 2(\lambda_{\max}(\mathbf{K})/\lambda_{\min}(\mathbf{K}) - 1)\|\hat{\boldsymbol{\alpha}}\|_\infty^2\mathbf{I}, \end{aligned}$$

where now the last inequality follows because the second to last expression is always non-negative.

Term 1 + 2 + 3: Putting together the inequalities for all terms, we get

$$\lambda_{\max}(\mathbf{H}_s) \leq 2\lambda_{\min}(\hat{\mathbf{K}})^{-1} - \lambda_{\max}(\hat{\mathbf{K}})^{-2} + 2(\lambda_{\max}(\mathbf{K})/\lambda_{\min}(\mathbf{K}) - 1)\|\hat{\boldsymbol{\alpha}}\|_\infty^2,$$

Finally, upper bounding the current upper bound by $M/\|(1-s)^{-1}\|_\infty^2 > 0$ yields a sufficient condition for the M -smoothness at s . Using the same bound for $\|\hat{\boldsymbol{\alpha}}\|_\infty$ derived for the convexity result,

$$2\lambda_{\min}(\hat{\mathbf{K}})^{-1} - \lambda_{\max}(\hat{\mathbf{K}})^{-2} + 2((\lambda_{\max}(\mathbf{K})/\lambda_{\min}(\mathbf{K}) - 1)\lambda_{\min}(\hat{\mathbf{K}})^{-2}\lambda_{\min}(\mathbf{K})^{-1}\|\mathbf{y}\|_2^2) < M.$$

Re-arranging, we attain

$$\lambda_{\min}\hat{\lambda}_{\min}^2 \frac{(M(1-s_{\max})^2 + \hat{\lambda}_{\min}^{-2} - 2\hat{\lambda}_{\max}^{-1})}{2(\lambda_{\max}/\lambda_{\min} - 1)} > \|\mathbf{y}\|_2^2, \quad (17)$$

which is a non-trivial guarantee when $M(1-s_{\max})^2 + \hat{\lambda}_{\min}^{-2} - 2\hat{\lambda}_{\max}^{-1} > 0$. \square

Lemma 7. [Strong Convexity via Diagonal Dominance] Let $m > 0$ and \mathbf{K}_0 be δ -diagonally dominant with $\delta < ((5-m) - \sqrt{25-9m+17})/4 \leq (5-\sqrt{17})/4 \approx 0.44$ and

$$\lambda_{\min}(\mathbf{K}_0)(1-\delta)^2 \frac{2(1+\delta)^{-1} - (1-\delta)^{-2} - m}{2(1-(1-\delta)/(1+\delta))} \geq \|\mathbf{y}\|_2^2.$$

Then the NMLL is m -strongly convex for all $s \in [0, 1]^n$, i.e. $\boldsymbol{\rho}(s) \in [0, \infty]^n$.

Proof. First, Gershgorin's Disk Theorem implies that the eigenvalues of $\operatorname{diag}(\mathbf{A})^{-1/2}\mathbf{A}\operatorname{diag}(\mathbf{A})^{-1/2}$ lie in $(1-\delta, 1+\delta)$ for a δ -diagonally dominant matrix \mathbf{A} . Further, the condition number of \mathbf{A} is bounded above by $\kappa(\mathbf{A}) = \lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A}) \leq (1+\delta)/(1-\delta)$. See Horn and Johnson [28] for more background on matrix analysis. Plugging these bounds into the results of Lemma 5 yields

$$\lambda_{\min}(\mathbf{K})\hat{\lambda}_{\min}^2 \frac{(2\hat{\lambda}_{\max}^{-1} - \hat{\lambda}_{\min}^{-2} - m)}{2(1-\lambda_{\min}/\lambda_{\max})} \geq \lambda_{\min}(\mathbf{K})(1-\delta)^2 \frac{2(1+\delta)^{-1} - (1-\delta)^{-2} - m}{2(1-(1-\delta)/(1+\delta))}.$$

Lower bounding the last expression by $\|\mathbf{y}\|_2^2$ implies m -strong convexity. This gives rise to a non-trivial guarantee whenever the numerator is larger than zero. In particular,

$$2(1+\delta)^{-1} - (1-\delta)^{-2} - m > 0.$$

Expanding, noting that $0 < \delta < 1$ implies $\delta^3 < \delta^2$ in order to reduce a power in the resulting expression, and collecting terms with like powers, we attain the following sufficient condition

$$2\delta^2 + (m-5)\delta + (1-m) > 0.$$

Note that for this condition to hold if $\delta = 0$, we need to have $m < 1$. Fortunately, this is a quadratic in δ whose smallest positive root is

$$\delta_- = \frac{1}{4} \left((5-m) - \sqrt{(5-m)^2 - 8(1-m)} \right).$$

In particular, for $m = 0$, this reduces to

$$\delta_- = \frac{1}{4} \left(5 - \sqrt{17} \right) \approx 0.4384471871911697.$$

Lastly, note that if \mathbf{K}_0 is δ -diagonally dominant, then so is $\mathbf{K}_s = \mathbf{K}_0 + \mathbf{D}_\rho(\mathbf{s})$, since the robust variances add to the diagonal, making it more dominant. Therefore, the convexity guarantee holds for all $\rho(\mathbf{s})$, if it holds for the base covariance matrix \mathbf{K}_0 . Note that $\lambda_{\min}(\mathbf{K}) \geq \sigma^2$. \square

Similarly, we can prove a similar statement relating diagonal dominance with M -smoothness.

Lemma 14. *[Smoothness via Diagonal Dominance] Suppose \mathbf{K}_0 is a δ -diagonally dominant covariance matrix and suppose we constrain $\|\mathbf{s}\|_\infty \leq s_{\max} \leq 1 - \sqrt{1/M}$. Then*

$$\lambda_{\min}(\mathbf{K}_0)(1 - \delta)^2 \frac{M(1 - s_{\max})^2 - 1}{2((1 + \delta)/(1 - \delta) - 1)} \geq \|\mathbf{y}\|_2^2,$$

implies that the NMLL is m -strongly convex for all $\mathbf{s} \in [0, 1]^n$, i.e. $\rho(\mathbf{s}) \in [0, \infty]^n$.

Proof. We proceed in a similar way as for Lemma 7, but with Lemma 13 as the starting point.

$$\lambda_{\min}(\mathbf{K})(\hat{\lambda}_{\min})^2 \frac{M(1 - s_{\max})^2 + \hat{\lambda}_{\min}^{-2} - 2\hat{\lambda}_{\max}^{-1}}{2(\lambda_{\max}/\lambda_{\min} - 1)} \geq \lambda_{\min}(\mathbf{K})(1 - \delta)^2 \frac{M(1 - s_{\max})^2 - 1}{2((1 + \delta)/(1 - \delta) - 1)},$$

where we used $(\hat{\lambda}_{\min}^{-2} - 2\hat{\lambda}_{\max}^{-1}) \geq -1$, which is tight when $\hat{\mathbf{K}}$ is unitary. Lower bounding the last expression by $\|\mathbf{y}\|_2^2$ implies M -smoothness. This gives rise to a non-trivial guarantee whenever the numerator is larger than zero. In particular, $M(1 - s_{\max})^2 - 1 > 0$, which implies $s_{\max} \leq 1 - \sqrt{1/M}$ or equivalently, $M > 1/(1 - s_{\max})^2$. Lastly, note that if \mathbf{K}_0 is δ -diagonally dominant, then so is $\mathbf{K}_s = \mathbf{K}_0 + \mathbf{D}_\rho(\mathbf{s})$, since the robust variances only add to the diagonal. Therefore, if the inequality holds for the base covariance matrix \mathbf{K}_0 , the smoothness guarantee holds for all $\rho(\mathbf{s})$ such that $\mathbf{s} \leq s_{\max} \leq 1 - \sqrt{1/M}$. Note also that $\lambda_{\min}(\mathbf{K}) \geq \sigma^2$. \square

Theorem 8. *[Approximation Guarantee] Let $\mathbf{K}_0 = k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}$ be δ -diagonally dominant, $s_{\max} > 0$ be an upper bound on $\|\mathbf{s}\|_\infty$, and suppose $\|\mathbf{y}\|, \delta$ satisfy the bounds of Lemmas 7 and 14, guaranteeing m -convexity and M -smoothness of the NMLL for some $m > 0$, $M > 1/(1 - s_{\max})^2$. Let $\mathbf{s}_{\text{OMP}}(r)$ be the r -sparse vector attained by OMP on the NMLL objective for r steps, and let $\mathbf{s}_{\text{OPT}}(r) = \arg \max_{\|\mathbf{s}\|_0=r, \|\mathbf{s}\|_\infty \leq s_{\max}} \mathcal{L}(\rho(\mathbf{s}))$ be the optimal r -sparse vector. Then for any $2r \leq n$,*

$$\tilde{\mathcal{L}}(\rho(\mathbf{s}_{\text{OMP}}(r))) \geq (1 - e^{-m/M}) \tilde{\mathcal{L}}(\rho(\mathbf{s}_{\text{OPT}}(r))),$$

where $\tilde{\mathcal{L}}(\cdot) = \mathcal{L}(\cdot) - \mathcal{L}(\mathbf{0})$ is normalized so that $\max_{\mathbf{s}_S} \tilde{\mathcal{L}}(\mathbf{s}_S) \geq 0$ for any support S .

Proof. The result is a direct consequence of meeting the m -convexity and M -smoothness conditions of Lemmas 7, 14 above, and the OMP approximation guarantee of Theorem 11 due to Elenberg et al. [21]. Note that the condition on $2r \leq n$ comes from the RSC condition in Theorem 11 being required for subsets of size $2r$. As we proved bounds for the m -convexity of the full Hessian of size n , r has to be smaller than $n/2$ for the assumptions of the theorem to hold. Regarding the upper bound s_{\max} on \mathbf{s} , we note that the constraint is convex and therefore doesn't change the convexity property of the optimization problem.

Further, note that $\max_{\rho_S} \mathcal{L}(\rho_S) \leq \max_{\rho_{S \cup i}} \mathcal{L}(\rho_{S \cup i})$, since the additional non-zero element could stay at 0, if the marginal likelihood does not improve with ρ_i increasing. That is, the subset selection problem is monotonic. As a consequence, we can normalize the MLL by $\tilde{\mathcal{L}}(\cdot) = \mathcal{L}(\cdot) - \mathcal{L}(\mathbf{0})$, which then only takes positive values for any $\mathbf{s}_S^* = \arg \max_{\mathbf{s}_S, \mathbf{s}_S=0} \mathcal{L}(\rho(\mathbf{s}_S))$, i.e. $\max_{\mathbf{s}_S} \tilde{\mathcal{L}}(\mathbf{s}_S) \geq 0$. This normalization is required for the constant factor approximation guarantee to apply, similar to the original work of Nemhauser. \square

This theoretical approach could lead to approximation guarantees for Tipping [55]'s Sparse Bayesian Learning (SBL) model, for which Ament and Gomes [8] show that greedily optimizing the associated NMLL is equivalent to stepwise regression in the limit of $\sigma \rightarrow 0$, proving exact recovery guarantees.

D Additional Detail on the Experiments

Our benchmark uses a modified version of the code from Andrade and Takeda [9], available at <https://github.com/andrade-stats/TrimmedMarginalLikelihoodGP> under the GNU GPLv2 license.

D.1 Synthetic Regression Problems

Model fitting runtimes Fig. 8 summarizes the fitting times for the different models on the different scenarios from Section 6.1. We observe that the outlier type and the fraction of outliers both have relatively limited effect on the fitting times for all of the models. Fig. 9 therefore provides a more compact view of the same data (aggregated across outlier types and outlier fractions), giving a better sense of the distribution of the fitting times. Unsurprisingly, the baselines that simply fit a single GP model (“vanilla”, “winsorize”, “power_transform”) are substantially faster than any of the robust approaches. While all of the robust models show similar fitting times on the Hartmann problem, fitting our RRP is significantly faster (note the logarithmic scale of the y-axis) than fitting the Student- t and trimmed MLE models on the 5-dim and 10-dim Friedman problems. The trimmed MLE model in particular ends up being quite slow, especially on the 5d Friedman function.

D.2 BO experiments, additional details

For Hartmann6, we consider the standard domain of $[0, 1]^6$.

SVM The SVM problem, the goal is to optimize the test RMSE of an SVM regression model trained on 100 features from the CT slice UCI dataset. We tune the following three parameters: $C \in [1e-2, 1e2]$, $\epsilon \in [1e-2, 1]$, and $\gamma \in [1e-3, 1e-1]$. All parameters are tuned in log-scale. Corruptions simulate I/O failures in which case we only train on $U[100, 1000]$ training points out of the available 50,000 training observations.

CNN For the 5D CNN problem the goal is to optimize the test accuracy of a CNN classifier trained on the MNIST dataset. We tune the following 5 parameters: learning rate in the interval $[1e-4, 1e-1]$, momentum in the interval $[0, 1]$, weight decay in the interval $[0, 1]$, step size in the interval $[1, 100]$, and $\gamma \in [0, 1]$. Similarly to the SVM problem, we only train on $U[100, 1000]$ of the available training batches when an I/O failure occurs.

Rover The rover trajectory planning problem was originally proposed in Wang et al. [60]. The goal is to tune the trajectory of a rover in order to maximize the reward of its final trajectory. We use the same obstacle locations and trajectory parameterization as in Maus et al. [42], with the main difference being that we parameterize the trajectory using 10 points, resulting in 20 tunable parameters. When a corruption occurs, the rover will stop at a uniformly random point along its trajectory, generally resulting in lower reward than the original trajectory.

D.3 Additional Synthetic BO Results

In addition to the results in Fig. 6, we also include results when the corruption probability is 20% in Fig. 10. We observe that the results are similar as with 10% corruption probability, but that the performance of several baselines regresses significantly.

D.4 Computational Setup and Requirements

Robust GP regression is very data-efficient, focuses on the small-data regime, and runs fast (faster than competing baselines studied in this paper). Therefore, each individual run required very limited compute resources (this includes the baseline methods). To produce statistically meaningful results, however, we ran a large number of replications for both our regression and Bayesian optimization benchmarks on a proprietary cluster. We estimate the amount of compute spent on these experiments to be around 2 CPU years in total, using standard (Intel Xeon) CPU hardware. The amount of compute spent on exploratory investigations as part of this work was negligible (this was ad-hoc exploratory and development work on a single CPU machine).

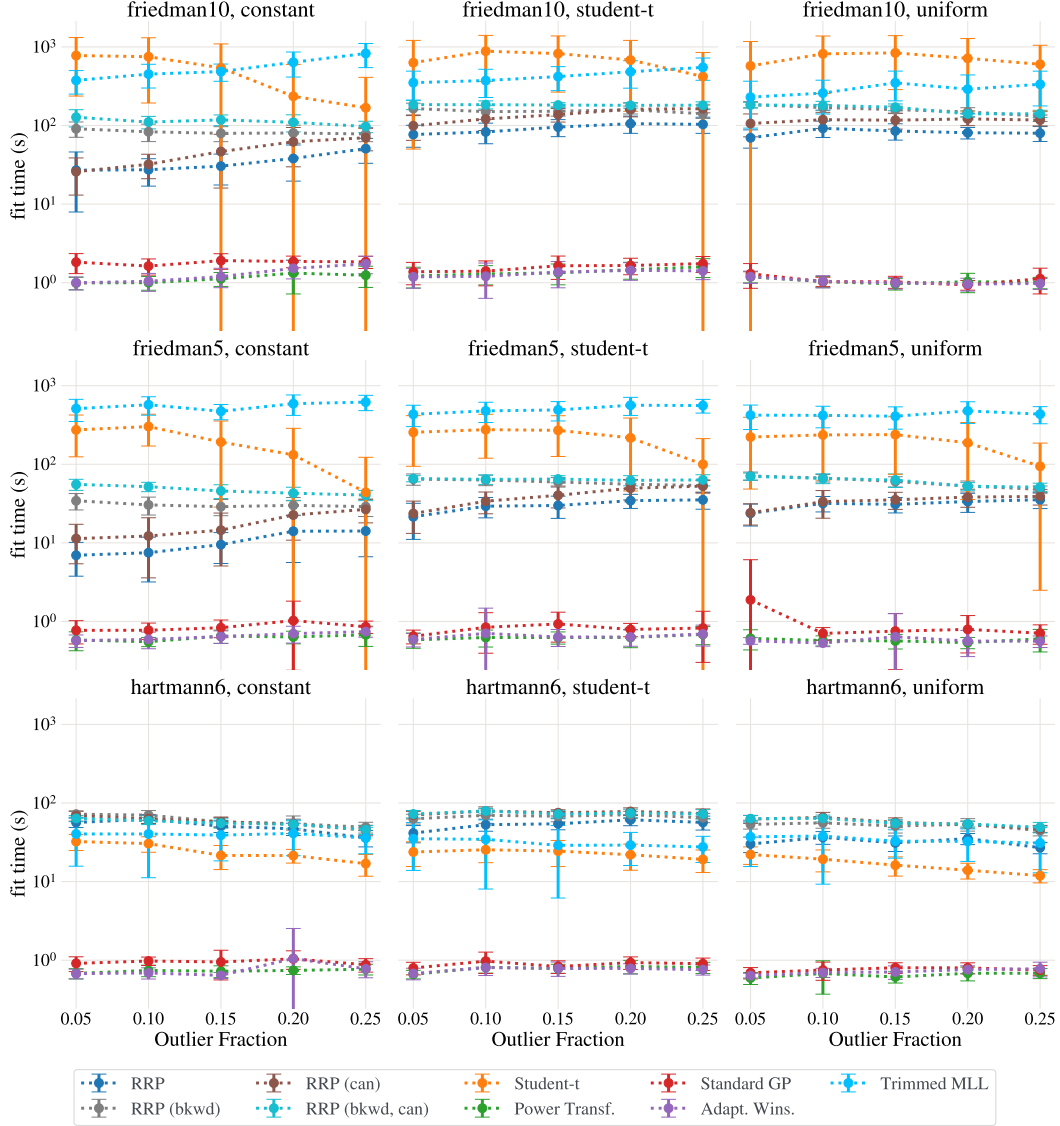


Figure 8: Fitting times of the different robust GP modeling approaches. The plots show means and one standard deviation. Here “bkwd” indicates the backward variant of RRP (Algorithm 2), and “can” indicates the canonical (non-convex) parameterization.

D.5 Impact of Convex Parameterization on Joint Optimization of Hyper-Parameters

A limitation of Lemma 4 is that it only guarantees convexity for the sub-problem of optimizing the ρ_i 's. In practice, we jointly optimize all GP hyper-parameters, including length scales. In this case, the theory guarantees the positive-definiteness of the *submatrix* of the Hessian corresponding to $\rho(\mathbf{s})$, and we expect this to improve the quality of the results of the numerical optimization routines.

Indeed, positive-definiteness is beneficial for quasi-Newton optimization algorithms like L-BFGS [38], which restarts its approximation to the Hessian whenever it encounters non-convex regions, because the associated updates to the Hessian approximation are not positive-definite. This leads the algorithm to momentarily revert back to gradient descent, with an associated slower convergence rate.

To quantify the impact of this, we ran convergence analyses using the data from Fig. 1, allowing all ρ to be optimized jointly with other hyper-parameters (length scale, kernel variance and noise

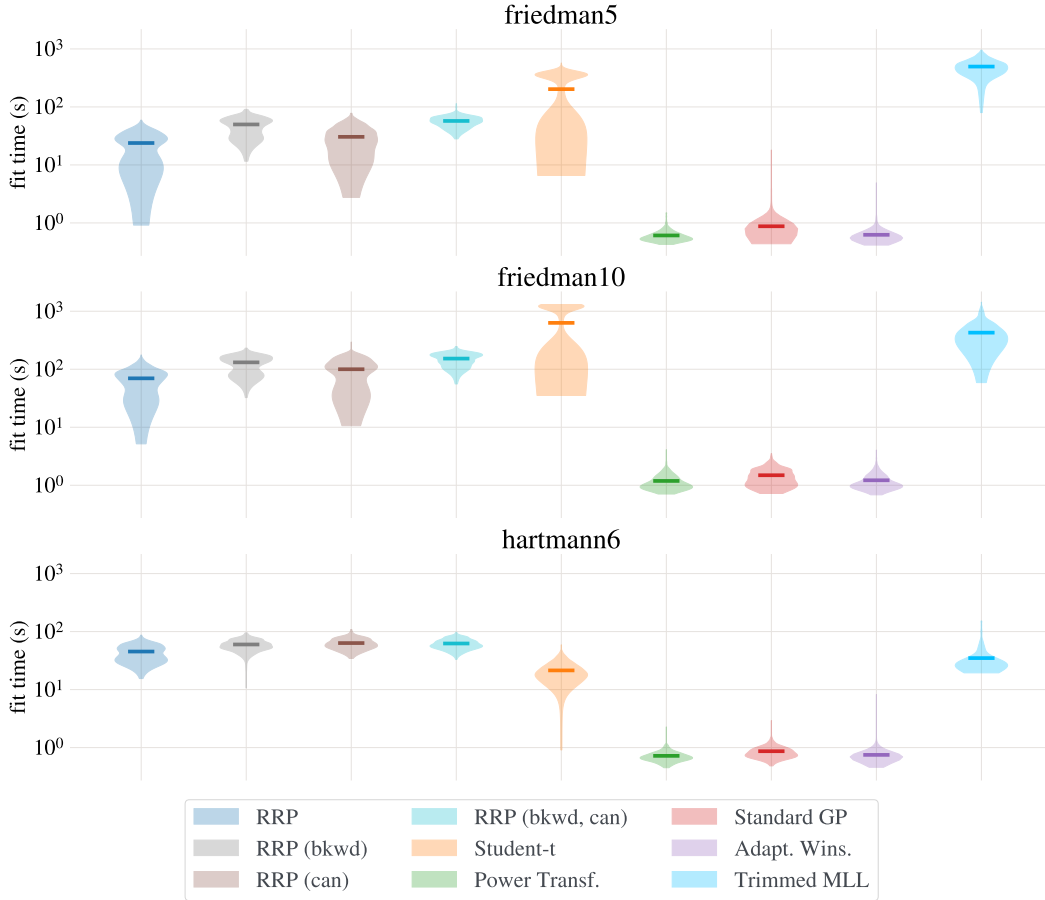


Figure 9: Fitting times of the different robust GP modeling approaches on the regression tasks from Section 6.1. Results are aggregated across outlier types and outlier fractions as those do not affect fitting times much (see Fig. 8).

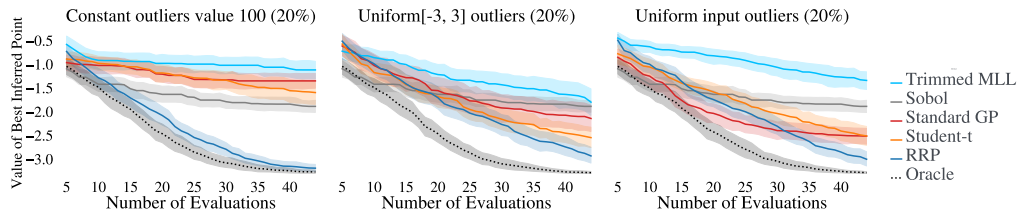


Figure 10: BO results for Hartmann6: *Left*: We see that Relevance pursuit performs well in the case of constant outliers of value 100 and almost performs just as well as the oracle. *Middle*: Relevance pursuit performs the best followed by the Student- t likelihood from [41] in the case of $U[-3, 3]$. No method performs as well as the oracle when the outlier probability is 20%. *Right*: Similarly to the middle column, this setting hides the outliers within the range of the outliers making it difficult to match the performance of the oracle.

variance), recording the achieved negative log marginal likelihood (NLML) as a function of the tolerance parameter ftol of the L-BFGS optimizer. The results are reported in Table D.5, and indicate that the optimizer terminates with a much better NLML using the convex parameterization with the same convergence tolerance.

There are also settings in which we do not actually jointly optimize the hyper-parameters, particularly when we have access to data from the same data generating process that has been manually labeled

ftol	Canonical ρ	Convex ρ (s)
1e-3	-4.37	-14.18
1e-4	-4.37	-93.00
1e-5	-4.37	-93.01
1e-6	-4.37	-135.05
1e-7	-98.62	-518.68
1e-8	-97.52	-1139.29

Table 1: Comparison of the negative log marginal likelihood achieved after numerical optimization of the canonical and convex parameterization of ρ with L-BFGS. Notably, the convex parameterization leads to improved marginal likelihood values for a given convergence tolerance ftol.

by domain experts as outlier-free. Then we can estimate the model hyper-parameters on that data, and fix them for the RRP on new data sets that we do not know to be outlier-free.

D.6 Comparison to Robust and Conjugate Gaussian Processes (RCGP)

We report extended empirical comparisons with Altamirano et al. [2]’s RCGP method, using their experimental setup and method implementation in GPFlow. Including GPFlow in our own benchmarking setup and compute resources proved difficult. To circumvent this, we wrote wrappers for both BoTorch’s standard GP and RRP, which also accounts for any orthogonal implementation differences between the two frameworks, and ran the benchmarks locally on an M-series MacBook.

See Tables 2 and 3 for the mean absolute error and negative log predictive density, respectively. The tables include the empirical mean and standard deviation over 20 replications on corrupted version of the following base data sets: 1) Synthetic, which is generated as a draw of a GP with a exponentiated quadratic kernel, and four data sets available on the UCI machine learning repository[32], in particular, 2) Boston [27], 3) Concrete [63], 4) Energy [58], and 5) Yacht [25]. The benchmark considers no corruptions (“No Outliers”), “Asymmetric Outliers”, which are uniform in x are shifted negatively in y , “Uniform Outliers”, which shift y in both directions (positively and negatively), and “Focused Outliers”, which form concentrated clusters in both x and y . Any bold entry in the table signifies a results that is within one standard-error of the best result’s one standard-error confidence bound.

Table 2: Mean absolute error (MAE) using Altamirano et al. [2]’s experimental setup in GPFlow. RRP is always competitive with the other methods, and outperforms them significantly for uniform and asymmetric outliers.

	Standard GP (GPFlow)	Student- t GP (GPFlow)	RCGP (GPFlow)	Standard GP (BoTorch)	RRP (BoTorch)
No Outliers					
Synthetic	8.82e-2 (2.12e-3)	8.81e-2 (2.09e-3)	8.77e-2 (2.08e-3)	8.90e-2 (2.19e-3)	8.90e-2 (2.19e-3)
Boston	2.24e-1 (7.11e-3)	2.24e-1 (6.50e-3)	2.24e-1 (7.29e-3)	2.08e-1 (5.53e-3)	2.08e-1 (5.75e-3)
Concrete	2.00e-1 (3.72e-3)	1.99e-1 (3.73e-3)	1.93e-1 (4.05e-3)	1.92e-1 (3.80e-3)	1.93e-1 (3.66e-3)
Energy	3.01e-2 (8.65e-4)	3.44e-2 (1.46e-3)	2.42e-2 (1.01e-3)	3.09e-2 (9.41e-4)	3.03e-2 (8.88e-4)
Yacht	1.69e-2 (2.09e-3)	2.02e-2 (1.29e-3)	2.19e-2 (4.63e-3)	1.49e-2 (1.95e-3)	1.42e-2 (1.58e-3)
Uniform Outliers					
Synthetic	3.45e-1 (1.48e-2)	2.93e-1 (1.23e-2)	2.15e-1 (9.30e-3)	3.48e-1 (1.62e-2)	8.99e-2 (2.37e-3)
Boston	4.97e-1 (2.80e-2)	3.86e-1 (1.52e-2)	4.94e-1 (1.81e-2)	6.81e-1 (2.25e-2)	2.83e-1 (3.16e-2)
Concrete	3.80e-1 (8.36e-3)	3.54e-1 (5.98e-3)	4.43e-1 (7.97e-3)	3.80e-1 (7.77e-3)	2.03e-1 (3.98e-3)
Energy	2.67e-1 (8.93e-3)	2.73e-1 (2.40e-2)	3.05e-1 (3.55e-2)	5.60e-1 (1.47e-2)	3.06e-2 (8.35e-4)
Yacht	3.23e-1 (1.68e-2)	2.29e-1 (1.54e-2)	4.18e-1 (3.48e-2)	6.81e-1 (3.21e-2)	1.20e-2 (1.04e-3)
Asymmetric Outliers					
Synthetic	1.11e+0 (2.24e-2)	1.02e+0 (2.34e-2)	7.72e-1 (3.31e-2)	1.17e+0 (2.55e-2)	2.71e-1 (9.96e-2)
Boston	6.92e-1 (1.40e-2)	5.80e-1 (1.24e-2)	5.31e-1 (2.33e-2)	8.55e-1 (3.04e-2)	3.35e-1 (4.33e-2)
Concrete	6.60e-1 (1.21e-2)	5.55e-1 (1.08e-2)	4.91e-1 (1.11e-2)	6.60e-1 (1.25e-2)	2.03e-1 (4.45e-3)
Energy	6.03e-1 (1.12e-2)	4.74e-1 (9.99e-3)	4.08e-1 (3.34e-2)	7.05e-1 (1.78e-2)	3.07e-2 (8.33e-4)
Yacht	5.90e-1 (1.60e-2)	4.58e-1 (9.31e-3)	4.32e-1 (3.30e-2)	8.36e-1 (3.76e-2)	2.70e-2 (1.35e-2)
Focused Outliers					
Synthetic	1.89e-1 (1.50e-2)	1.92e-1 (1.56e-2)	1.64e-1 (1.29e-2)	1.80e-1 (1.46e-2)	2.00e-1 (1.62e-2)
Boston	2.44e-1 (8.48e-3)	2.60e-1 (1.30e-2)	2.49e-1 (1.04e-2)	2.49e-1 (8.37e-3)	2.48e-1 (8.38e-3)
Concrete	2.40e-1 (5.54e-3)	2.35e-1 (5.63e-3)	2.38e-1 (1.10e-2)	2.24e-1 (6.20e-3)	2.25e-1 (5.86e-3)
Energy	8.83e-2 (5.46e-2)	3.24e-2 (1.36e-3)	2.92e-2 (8.95e-4)	3.03e-2 (8.95e-4)	3.03e-2 (8.95e-4)
Yacht	2.42e-1 (6.89e-2)	1.10e-1 (2.07e-2)	1.93e-2 (1.88e-3)	1.78e-2 (2.96e-3)	1.59e-2 (2.20e-3)

D.7 Comparison to Robust Gaussian Process with Huber Likelihood

In the following, we compare our method with additional variational GP baselines with Laplace and Huber likelihoods, and translated the Matlab code of the "projection statistics" of Algikar and Mili [1] to PyTorch. We then combined the projection-statistics-based weighting of the Huber loss with

Table 3: Negative log predictive density (NLPD) using Altamirano et al. [2]’s experimental setup in GPFlow. RRP is generally competitive, and outperforms other methods significantly for uniform and asymmetric outliers.

	Standard GP (GPFlow)	Student- <i>t</i> GP (GPFlow)	RCGP (GPFlow)	Standard GP (BoTorch)	RRP (BoTorch)
No Outliers					
Synthetic	-8.21e-1 (2.16e-2)	-8.20e-1 (2.16e-2)	-8.23e-1 (2.21e-2)	-8.05e-1 (2.25e-2)	-8.05e-1 (2.24e-2)
Boston	2.08e-1 (4.26e-2)	1.99e-1 (3.76e-2)	1.95e-1 (4.69e-2)	9.24e-2 (2.82e-2)	9.43e-2 (2.82e-2)
Concrete	1.48e-1 (2.33e-2)	1.29e-1 (2.31e-2)	1.09e-1 (3.19e-2)	1.11e-1 (2.73e-2)	1.16e-1 (2.68e-2)
Energy	-1.72e+0 (3.83e-2)	-1.62e+0 (3.46e-2)	-1.96e+0 (4.37e-2)	-1.67e+0 (3.08e-2)	-1.69e+0 (4.02e-2)
Yacht	-1.79e+0 (3.23e-1)	-2.05e+0 (6.61e-2)	-2.00e+0 (3.27e-1)	-2.47e+0 (1.06e-1)	-2.23e+0 (1.84e-1)
Uniform Outliers					
Synthetic	1.60e+0 (1.41e-2)	1.47e+0 (1.55e-2)	1.57e+0 (1.39e-2)	1.58e+0 (1.40e-2)	-7.99e-1 (2.26e-2)
Boston	1.67e+0 (7.55e-3)	1.54e+0 (8.98e-3)	1.64e+0 (1.12e-2)	1.81e+0 (5.95e-2)	4.52e-1 (1.34e-1)
Concrete	1.64e+0 (6.83e-3)	1.51e+0 (6.56e-3)	1.63e+0 (7.57e-3)	1.63e+0 (6.72e-3)	1.47e-1 (2.51e-2)
Energy	1.61e+0 (7.75e-3)	1.50e+0 (1.76e-2)	1.61e+0 (9.74e-3)	1.70e+0 (8.58e-3)	-1.67e+0 (3.72e-2)
Yacht	1.61e+0 (9.50e-3)	1.46e+0 (1.19e-2)	1.62e+0 (1.87e-2)	1.78e+0 (4.67e-2)	-2.32e+0 (1.97e-1)
Asymmetric Outliers					
Synthetic	1.94e+0 (9.39e-3)	1.90e+0 (9.98e-3)	1.89e+0 (1.13e-2)	1.93e+0 (1.06e-2)	-3.70e-1 (2.30e-1)
Boston	1.68e+0 (1.01e-2)	1.56e+0 (1.09e-2)	1.64e+0 (9.51e-3)	2.52e+0 (7.26e-1)	6.14e-1 (1.19e-1)
Concrete	1.66e+0 (7.43e-3)	1.54e+0 (7.45e-3)	1.62e+0 (6.63e-3)	1.65e+0 (7.55e-3)	1.55e-1 (2.86e-2)
Energy	1.62e+0 (8.85e-3)	1.49e+0 (9.70e-3)	1.62e+0 (1.82e-2)	1.71e+0 (9.79e-3)	-1.66e+0 (3.55e-2)
Yacht	1.59e+0 (9.35e-3)	1.45e+0 (1.01e-2)	1.56e+0 (8.15e-3)	1.75e+0 (1.41e-2)	-2.17e-1 (1.82e+0)
Focused Outliers					
Synthetic	6.78e-1 (6.59e-2)	6.11e-1 (8.68e-2)	5.69e-1 (4.07e-2)	6.66e-1 (6.60e-2)	9.67e+0 (2.24e+0)
Boston	2.57e-1 (4.23e-2)	3.21e-1 (6.40e-2)	2.74e-1 (6.13e-2)	1.98e-1 (3.04e-2)	1.97e-1 (3.06e-2)
Concrete	2.77e-1 (2.58e-2)	2.72e-1 (2.81e-2)	2.43e-1 (4.26e-2)	2.22e-1 (2.87e-2)	2.28e-1 (2.72e-2)
Energy	2.81e+4 (2.81e+4)	-1.67e+0 (4.13e-2)	-1.74e+0 (4.98e-2)	-1.70e+0 (4.59e-2)	-1.70e+0 (4.59e-2)
Yacht	9.38e+4 (5.09e+4)	-4.60e-1 (2.91e-1)	-2.47e+0 (8.09e-2)	-2.63e+0 (5.55e-2)	-2.61e+0 (5.66e-2)

a variational (referred to as Huber-Projection) to get as close as possible to a direct comparison to Algikar and Mili [1] without access to a Matlab license.

Tables 4 and 5 shows the root mean square error and negative log predictive density on the Neal, Friedman 5 and Friedman 10 test functions, as well as the Yacht Hydrodynamics [25] and California Housing [46] datasets from the UCI database [32], where 15% of the training data sets of the models were corrupted. Tables 6 and 7 below were generated in a similar way, but 100% of the data were subject to heavier-tailed noise, either Student-*t* or Laplace.

In summary, the Relevance Pursuit model generally outperforms the variational GPs with heavy-tailed likelihoods when the corruptions are a sparse subset of all observations. Unsurprisingly, the GPs with heavy-tailed likelihoods perform best when *all* observations are subject to heavy-tailed noise. While such uniformly heavy-tailed noise does exist in practice, we stress that this is a distinct setting to the common setting where datasets contain a *subset* of a-priori unknown outliers, while a dominant fraction of the data can be considered inliers that, once they are identified, can be used to train a model without additional treatment.

Table 4: Comparison with Huber GP: Root mean square error with 15% Corruptions

Data	Standard	Relevance Pursuit	Student- t	Laplace	Huber	Huber + Projection
Neal						
Uniform	3.87e-1 (2.40e-2)	3.79e-2 (1.10e-2)	4.18e-1 (8.65e-3)	4.18e-1 (4.72e-3)	4.18e-1 (4.76e-3)	4.18e-1 (4.76e-3)
Constant	1.91e+0 (1.18e-1)	4.37e-2 (1.87e-2)	7.26e-1 (1.38e-1)	4.73e-1 (1.28e-2)	4.89e-1 (1.62e-2)	4.89e-1 (1.62e-2)
Student- t	7.74e-1 (2.70e-1)	6.71e-2 (2.29e-2)	3.90e-1 (3.59e-3)	4.07e-1 (4.79e-3)	4.09e-1 (5.02e-3)	4.09e-1 (5.02e-3)
Laplace	7.19e-1 (6.17e-2)	6.27e-2 (1.52e-2)	4.01e-1 (4.42e-3)	4.20e-1 (4.34e-3)	4.20e-1 (4.00e-3)	4.20e-1 (4.00e-3)
Friedman 5						
Uniform	2.05e+0 (8.83e-2)	9.40e-2 (1.23e-2)	7.01e-1 (1.37e-1)	7.30e-1 (7.33e-2)	6.78e-1 (6.67e-2)	7.12e-1 (7.50e-2)
Constant	1.48e+1 (5.16e-1)	7.29e-2 (5.40e-3)	8.98e-1 (1.40e-1)	2.37e+0 (5.19e-2)	2.50e+0 (7.18e-2)	2.50e+0 (7.18e-2)
Student- t	7.91e+0 (8.80e-1)	1.17e-1 (2.24e-2)	5.26e-1 (1.26e-1)	1.27e+0 (9.49e-2)	1.36e+0 (9.34e-2)	1.31e+0 (9.73e-2)
Laplace	7.59e+0 (4.98e-1)	9.63e-2 (1.15e-2)	6.62e-1 (1.34e-1)	1.51e+0 (7.78e-2)	1.63e+0 (6.89e-2)	1.63e+0 (6.89e-2)
Friedman 10						
Uniform	1.67e+0 (4.94e-2)	4.84e-2 (6.21e-3)	5.92e-1 (1.36e-1)	6.82e-1 (1.03e-1)	6.74e-1 (1.04e-1)	5.93e-1 (9.43e-2)
Constant	1.34e+1 (2.97e-1)	5.12e-2 (2.69e-3)	8.63e-1 (1.39e-1)	2.10e+0 (2.39e-2)	2.18e+0 (3.01e-2)	2.18e+0 (3.01e-2)
Student- t	7.93e+0 (1.75e+0)	8.45e-2 (2.42e-2)	5.30e-1 (1.30e-1)	1.40e+0 (9.10e-2)	1.40e+0 (9.12e-2)	1.40e+0 (9.04e-2)
Laplace	6.42e+0 (3.42e-1)	5.70e-2 (8.16e-3)	7.23e-1 (1.39e-1)	1.64e+0 (4.98e-2)	1.72e+0 (1.06e-2)	1.72e+0 (1.06e-2)
Yacht						
Uniform	6.72e+0 (5.08e-1)	8.56e-1 (4.99e-2)	1.62e+1 (1.81e-1)	1.62e+1 (2.08e-1)	1.61e+1 (2.04e-1)	1.61e+1 (1.94e-1)
Constant	3.50e+1 (1.61e+0)	8.27e-1 (5.78e-2)	1.60e+1 (2.11e-1)	1.59e+1 (2.02e-1)	1.55e+1 (1.73e-1)	1.58e+1 (1.95e-1)
Student- t	2.11e+1 (3.51e+0)	1.12e+0 (2.03e-1)	1.64e+1 (2.17e-1)	1.67e+1 (2.15e-1)	1.64e+1 (2.08e-1)	1.66e+1 (2.02e-1)
Laplace	1.61e+1 (1.97e+0)	1.15e+0 (1.84e-1)	1.63e+1 (2.10e-1)	1.64e+1 (4.06e-1)	1.63e+1 (2.05e-1)	1.66e+1 (1.99e-1)
CA Housing						
Uniform	7.10e-1 (7.58e-3)	7.39e-1 (1.75e-2)	1.16e+0 (1.79e-3)	1.17e+0 (3.04e-3)	1.17e+0 (2.93e-3)	1.18e+0 (6.17e-3)
Constant	2.28e+0 (5.12e-2)	6.35e-1 (4.38e-3)	1.17e+0 (2.39e-3)	1.16e+0 (1.87e-3)	1.16e+0 (1.88e-3)	1.17e+0 (5.51e-3)
Student- t	1.34e+0 (1.68e-1)	6.56e-1 (5.46e-3)	1.18e+0 (3.56e-3)	1.19e+0 (4.13e-3)	1.18e+0 (3.83e-3)	1.19e+0 (6.74e-3)
Laplace	1.00e+0 (5.04e-2)	6.51e-1 (4.71e-3)	1.18e+0 (3.41e-3)	1.18e+0 (3.91e-3)	1.18e+0 (3.67e-3)	1.18e+0 (6.30e-3)

Table 5: Comparison with Huber GP: Negative log predictive density with 15% Corruptions

Data	Standard	Relevance Pursuit	Student- t	Laplace	Huber	Huber + Projection
Neal						
Uniform	6.87e+0 (4.15e+0)	3.80e+0 (4.01e+0)	1.85e+0 (9.93e-2)	1.64e+0 (1.20e-1)	1.70e+0 (1.23e-1)	1.70e+0 (1.23e-1)
Constant	1.32e+2 (1.29e+2)	-2.50e+0 (2.99e-1)	2.03e+0 (1.84e-1)	8.67e-1 (4.32e-2)	8.84e-1 (3.87e-2)	8.84e-1 (3.87e-2)
Student- t	1.98e+0 (9.67e-1)	1.99e-1 (2.29e+0)	1.67e+0 (1.17e-1)	1.25e+0 (8.32e-2)	1.28e+0 (8.03e-2)	1.28e+0 (8.03e-2)
Laplace	3.07e+2 (2.15e+2)	2.02e+0 (2.83e+0)	1.73e+0 (1.10e-1)	1.20e+0 (8.89e-2)	1.21e+0 (8.99e-2)	1.21e+0 (8.99e-2)
Friedman 5						
Uniform	2.07e+0 (4.88e-2)	-1.35e+0 (7.22e-2)	4.67e-1 (3.00e-1)	8.88e-1 (1.01e-1)	8.29e-1 (9.14e-2)	8.88e-1 (1.08e-1)
Constant	4.28e+0 (2.25e-2)	-1.13e+0 (1.82e-2)	8.24e-1 (3.17e-1)	2.28e+0 (2.63e-2)	2.34e+0 (3.13e-2)	2.34e+0 (3.13e-2)
Student- t	3.42e+0 (1.34e-1)	-1.11e+0 (1.04e-1)	1.78e-2 (2.86e-1)	1.55e+0 (1.14e-1)	1.66e+0 (1.11e-1)	1.59e+0 (1.10e-1)
Laplace	3.55e+0 (7.84e-2)	-1.22e+0 (5.89e-2)	3.47e-1 (3.13e-1)	1.75e+0 (8.46e-2)	1.90e+0 (7.36e-2)	1.90e+0 (7.36e-2)
Friedman 10						
Uniform	1.86e+0 (3.04e-2)	-1.82e+0 (3.97e-2)	6.78e-2 (3.85e-1)	9.20e-1 (1.99e-1)	9.39e-1 (2.01e-1)	7.65e-1 (1.80e-1)
Constant	4.21e+0 (1.39e-2)	-1.27e+0 (1.10e-2)	8.85e-1 (3.79e-1)	2.20e+0 (1.45e-2)	2.23e+0 (1.54e-2)	2.23e+0 (1.54e-2)
Student- t	3.34e+0 (1.24e-1)	-1.40e+0 (1.17e-1)	-6.80e-2 (3.65e-1)	1.96e+0 (1.25e-1)	1.95e+0 (1.23e-1)	1.96e+0 (1.22e-1)
Laplace	3.46e+0 (6.68e-2)	-1.44e+0 (5.91e-2)	4.89e-1 (3.94e-1)	2.13e+0 (6.35e-2)	2.21e+0 (2.01e-2)	2.21e+0 (2.01e-2)
Yacht						
Uniform	3.45e+0 (1.42e-1)	2.37e+0 (2.79e-1)	1.18e+2 (1.06e+1)	7.07e+1 (5.17e+0)	7.46e+1 (5.38e+0)	1.89e+2 (1.64e+1)
Constant	5.29e+0 (1.70e-1)	1.79e+0 (2.43e-1)	7.16e+1 (7.68e+0)	2.32e+1 (1.59e+0)	1.95e+1 (1.12e+0)	3.89e+1 (3.46e+0)
Student- t	4.46e+0 (3.08e-1)	2.42e+0 (3.38e-1)	1.32e+2 (1.05e+1)	8.20e+1 (5.07e+0)	8.13e+1 (5.17e+0)	1.55e+2 (1.38e+1)
Laplace	4.35e+0 (2.88e-1)	2.27e+0 (3.71e-1)	1.14e+2 (9.86e+0)	6.85e+1 (4.72e+0)	6.12e+1 (3.87e+0)	1.17e+2 (8.89e+0)
CA Housing						
Uniform	3.14e+0 (1.44e-1)	2.92e+0 (2.06e-1)	5.85e+1 (9.59e-1)	6.77e+1 (2.41e+0)	6.69e+1 (1.94e+0)	8.94e+1 (2.45e+1)
Constant	3.57e+0 (2.26e-1)	4.51e+0 (2.11e-1)	4.02e+1 (1.59e+0)	1.83e+1 (6.58e-1)	1.75e+1 (7.03e-1)	2.18e+1 (4.19e+0)
Student- t	1.77e+0 (6.65e-2)	3.15e+0 (1.67e-1)	5.95e+1 (1.44e+0)	5.20e+1 (2.19e+0)	4.95e+1 (1.75e+0)	6.90e+1 (2.02e+1)
Laplace	1.61e+0 (4.21e-2)	3.51e+0 (1.93e-1)	5.60e+1 (1.58e+0)	4.23e+1 (1.72e+0)	4.06e+1 (1.47e+0)	5.62e+1 (1.66e+1)

Table 6: Comparison with Huber GP: Root mean squared error for 100% Laplace noise

Data	Standard	Relevance Pursuit	Student- t	Laplace	Huber	Huber + Projection
Neal						
Student- t	1.96e+0 (3.64e-1)	1.56e+0 (2.48e-1)	7.39e-1 (4.08e-2)	7.54e-1 (3.17e-2)	7.48e-1 (3.00e-2)	7.48e-1 (3.01e-2)
Laplace	1.51e+0 (1.06e-1)	2.40e+0 (2.72e-1)	1.16e+0 (9.18e-2)	1.12e+0 (9.90e-2)	1.12e+0 (9.80e-2)	1.12e+0 (9.80e-2)
Friedman 5						
Student- t	1.67e+1 (1.83e+0)	9.30e+0 (3.05e-1)	5.26e+0 (1.30e-1)	4.57e+0 (1.60e-1)	4.61e+0 (1.55e-1)	4.61e+0 (1.55e-1)
Laplace	1.39e+1 (5.95e-1)	1.37e+1 (6.36e-1)	8.33e+0 (3.69e-1)	7.34e+0 (3.50e-1)	7.40e+0 (3.45e-1)	7.40e+0 (3.45e-1)
Friedman 10						
Student- t	2.10e+1 (3.42e+0)	7.80e+0 (2.03e-1)	4.72e+0 (1.14e-1)	4.02e+0 (9.57e-2)	4.05e+0 (9.34e-2)	4.05e+0 (9.34e-2)
Laplace	1.30e+1 (3.77e-1)	1.27e+1 (3.99e-1)	7.24e+0 (1.99e-1)	6.07e+0 (1.91e-1)	6.26e+0 (1.71e-1)	6.26e+0 (1.71e-1)
Yacht						
Student- t	5.89e+1 (1.37e+1)	2.44e+1 (1.77e+0)	1.57e+1 (1.36e-1)	1.57e+1 (1.51e-1)	1.57e+1 (1.51e-1)	1.60e+1 (3.16e-1)
Laplace	2.57e+1 (1.17e+0)	4.75e+1 (3.51e+0)	1.64e+1 (3.18e-1)	1.61e+1 (2.52e-1)	1.62e+1 (2.60e-1)	1.67e+1 (3.55e-1)
CA Housing						
Student- t	2.92e+0 (5.80e-1)	1.24e+0 (6.17e-2)	1.17e+0 (3.65e-3)	1.18e+0 (4.42e-3)	1.17e+0 (4.14e-3)	1.21e+0 (3.05e-2)
Laplace	1.57e+0 (8.25e-2)	2.06e+0 (1.44e-1)	1.21e+0 (1.08e-2)	1.20e+0 (9.39e-3)	1.20e+0 (9.69e-3)	1.23e+0 (2.05e-2)

Table 7: Comparison with Huber GP: Negative log predictive density with 100% Laplace noise

Data	Standard	Relevance Pursuit	Student- <i>t</i>	Laplace	Huber	Huber + Projection
Neal						
Student- <i>t</i>	1.95e+0 (1.22e-1)	8.78e+1 (2.80e+1)	1.17e+0 (6.08e-2)	1.22e+0 (5.71e-2)	1.20e+0 (5.17e-2)	1.21e+0 (5.15e-2)
Laplace	1.89e+0 (1.14e-1)	1.38e+2 (3.98e+1)	1.57e+0 (9.22e-2)	1.55e+0 (9.20e-2)	1.55e+0 (9.12e-2)	1.55e+0 (9.12e-2)
Friedman 5						
Student- <i>t</i>	4.43e+0 (8.19e-2)	3.93e+0 (2.19e-2)	3.11e+0 (2.69e-2)	2.96e+0 (3.49e-2)	2.97e+0 (3.27e-2)	2.97e+0 (3.27e-2)
Laplace	4.55e+0 (2.81e-2)	4.56e+0 (1.88e-2)	3.64e+0 (4.03e-2)	3.46e+0 (4.45e-2)	3.48e+0 (4.31e-2)	3.48e+0 (4.31e-2)
Friedman 10						
Student- <i>t</i>	4.60e+0 (8.54e-2)	3.89e+0 (1.36e-2)	3.04e+0 (2.45e-2)	2.83e+0 (2.30e-2)	2.84e+0 (2.22e-2)	2.84e+0 (2.22e-2)
Laplace	4.58e+0 (1.25e-2)	4.57e+0 (1.27e-2)	3.60e+0 (2.53e-2)	3.32e+0 (3.01e-2)	3.37e+0 (2.61e-2)	3.37e+0 (2.61e-2)
Yacht						
Student- <i>t</i>	5.14e+0 (1.85e-1)	4.60e+0 (8.15e-2)	7.30e+0 (1.40e-1)	6.79e+0 (1.52e-1)	6.70e+0 (1.32e-1)	1.28e+1 (7.75e-1)
Laplace	4.74e+0 (4.45e-2)	5.30e+0 (8.31e-2)	4.91e+0 (8.87e-2)	5.19e+0 (1.17e-1)	5.12e+0 (1.04e-1)	9.90e+0 (6.44e-1)
CA Housing						
Student- <i>t</i>	2.08e+0 (1.01e-1)	1.79e+0 (6.40e-2)	7.41e+0 (2.39e-1)	6.52e+0 (2.54e-1)	6.41e+0 (2.07e-1)	6.83e+0 (3.28e-1)
Laplace	1.88e+0 (5.25e-2)	2.24e+0 (9.41e-2)	2.92e+0 (6.01e-2)	3.33e+0 (7.47e-2)	3.27e+0 (5.84e-2)	6.64e+0 (3.28e+0)

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Everything that is claimed in the abstract and introduction is discussed in detail in the rest of the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations of the work are mentioned throughout the paper as well as summarized in the conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results clearly state the required assumptions; proofs are provided in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Code for our method as well as for the benchmarking setup is provided in the supplementary material, and core models and algorithms will be open-sourced upon publication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code for reproducing the results in the paper is included in the supplementary material. All data sets are publicly available and referenced in the work and the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Details on the experiments are provided in the experimental section as well as the Appendix. Exact implementation details are contained in the included code submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Variance in the results is provided throughout in the form of confidence intervals and/or violin plots.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Total compute resources spent are described in Section D.4 and are estimated at around 2 CPU years in total.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics [https://neurips.cc/public/EthicsGuidelines?](https://neurips.cc/public/EthicsGuidelines)

Answer: [Yes]

Justification: The research fully conforms to the NeurIPS Code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper makes foundational methodological contributions to robust probabilistic regression, and we do not see any direct societal impacts of the work that would require specific discussion.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All creators or original owners of the code and data sets that were used in this paper are properly cited and credited. Licenses for these assets are mentioned in Sec. D and our use complies with those licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The only new asset introduced in this paper is the code for method and benchmarks; it does not introduce any other assets (data sets or models). The code is provided as an anonymized zip file for the reviewers. Upon publication, the code released as open source and documented according to community standards.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.