Robust Gaussian Process Regression with the Trimmed Marginal Likelihood (Supplementary Material)

Daniel Andrade[1] **Akiko Takeda**[2,3]

[1]Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Hiroshima, Japan
[2]Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan
[3]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

A PROOFS

A.1 CONVERGENCE GUARANTEE OF THE PROPOSED PROJECTED GRADIENT DESCENT METHOD

Optimization problem (P1) from the main paper is given by

$$\min_{\mathbf{b}} f(\mathbf{b}) \text{ s.t. } \|\mathbf{b}\|_0 = n - m. \tag{1}$$

This problem can be expressed as an unconstrained optimization problem by using the indicator function[1] as follows:

$$\min_{\mathbf{b}} F(\mathbf{b}) \,, \tag{2}$$

with

$$F(\mathbf{b}) := f(\mathbf{b}) + \delta_C(\mathbf{b}), \text{ where } C = \{\mathbf{b} \in \mathbb{R}^n \mid \|\mathbf{b}\|_0 = n - m\} \,.$$

Recently, for analyzing the convergence rate of first-order methods for nonconvex objective functions, the so-called Kurdyka–Lojasiewicz (KL) property is often used. If the objective function of $F(\mathbf{b})$ satisfies the KL property with an exponent of $\alpha = 1/2$ and the sequence $\{b_k\}$ generated by the proximal gradient algorithm is bounded, then it was proven that $\{b_k\}$ converges locally and linearly to a stationary point of F (see, for example, Attouch et al. [2010, 2013], Li and Pong [2018]). Therefore, here, we only need to prove that $F(\mathbf{b})$ is a KL function with exponent $1/2$.

The definition of KL functions encompasses broad classes of functions, and it is known that a proper closed semi-algebraic function is a KL function with a suitable exponent $\alpha \in [0, 1)$. The above function F is also a KL function.

Theorem 1. *Any sequence $\{b_k\}$ generated by projected gradient algorithm for Problem* (1) *globally converges to a stationary point with locally linear convergence rate.*

Proof. First, we show global convergence. Bolte et al. [2014] implies that the objective function F of (2) is a proper lower semi-continuous KL function. Considering that F is lower bounded and ∇f is Lipschitz continuous, we can confirm the global convergence of the proximal gradient method from [Attouch et al., 2013, Theorem 5.1 and Remark 5.2]. Now for proving the convergence rate, we will check the KL exponent of F. F can be further rewritten as

$$F(\mathbf{b}) = \min_{S \subseteq \{1,\ldots,n\}, |S| = m} f(\mathbf{b}) + \delta_{\Omega_S}(\mathbf{b}) \,,$$

where $\Omega_S := \{\mathbf{b} \in \mathbb{R}^n \mid b_i = 0, \forall i \in S\}$. Here, for all possible S, $\delta_{\Omega_S}(\mathbf{b})$ are proper closed polyhedral functions. Then [Li and Pong, 2018, Corollary 5.2] implies that $F(\mathbf{b})$ is a KL function with an exponent of $1/2$. From this, and the boundedness of $\{b_k\}$, [Li and Pong, 2018, Proposition 5.1] implies that $\{b_k\}$ achieves linear convergence locally. $\square$

[1]The indicator function is defined as $\delta_C(\mathbf{b}) := \begin{cases} 0 & \text{if } \mathbf{b} \in C \,, \\ \infty & \text{else.} \end{cases}$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

A.2 PROOF OF ASYMPTOTICALLY CORRECT OUTLIER REJECTION

Here we prove Proposition 1. Note that ignoring constants, we may write the negative marginal log-likelihood (NLL) as

$$\begin{aligned}\mathrm{NLL}(\sigma^2, \eta, \mathbf{l}) &:= -2\log p(\mathbf{y}|X, \sigma^2, \eta, \mathbf{l}) - n\log 2\pi \\ &= \mathbf{y}^T(K_{\eta,\mathbf{l}} + \sigma^2 I)^{-1}\mathbf{y} + \log|K_{\eta,\mathbf{l}} + \sigma^2 I| \\ &= \frac{1}{\eta}\mathbf{y}^T(K + \frac{\sigma^2}{\eta}I)^{-1}\mathbf{y} + \log(\eta^n|K + \frac{\sigma^2}{\eta}I|)\,,\end{aligned}$$

where $K := K_{1,\mathbf{l}}$ (that means K is $K_{\eta,\mathbf{l}}$, with η being set to 1).

First, we establish a lower bound on NLL. Let λ_0 denote the smallest possible eigenvalue of $K_{1,\mathbf{l}}$, i.e.

$$\lambda_0 := \min_{\mathbf{l}\in\mathbb{D}} \lambda_{\min}(K_{1,\mathbf{l}})\,,$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a matrix A. Note that $1 \geq \lambda_0 > 0$. Analogously, let λ_1 denote the largest possible eigenvalue of $K_{1,\mathbf{l}}$, i.e.

$$\lambda_1 := \min_{\mathbf{l}\in\mathbb{D}} \lambda_{\max}(K_{1,\mathbf{l}})\,,$$

where $\lambda_{\max}(A)$ denotes the largest eigenvalue of a matrix A. Note that $1 \leq \lambda_1 < n$. Therefore, for any $\mathbf{l} \in \mathbb{D}$, all eigenvalues of K are bounded. In particular, we have

$$\lambda_{\min}\Big(K + \frac{\sigma^2}{\eta}\Big) \geq \lambda_0 + \frac{\sigma^2}{\eta}\,,$$

and

$$\lambda_{\min}\Big((K + \frac{\sigma^2}{\eta})^{-1}\Big) \geq (\lambda_1 + \frac{\sigma^2}{\eta})^{-1}\,.$$

Define

$$g_2(\sigma^2, \eta) := \frac{1}{\eta}(\lambda_1 + \frac{\sigma^2}{\eta})^{-1}||\mathbf{y}||_2^2 + \log(\eta^n(\lambda_0 + \frac{\sigma^2}{\eta})^n)\,,$$

then we have

$$g_2(\sigma^2, \eta) \leq \mathrm{NLL}(\sigma^2, \eta, \mathbf{l})\,.$$

Since the function g_2 is still slightly difficult to analyze, we establish another lower bounding function g_1.

First note that g_2 can be written as follows

$$g_2(\sigma^2, \eta) = (\eta\lambda_1 + \sigma^2)^{-1}||\mathbf{y}||_2^2 + n\log(\eta\lambda_0 + \sigma^2)\,.$$

Noting that

$$\begin{aligned}n\log(\lambda_0) + n\log(\eta + \sigma^2) &= n\log(\lambda_0\eta + \lambda_0\sigma^2) \\ &\leq n\log(\eta\lambda_0 + \sigma^2)\,,\end{aligned}$$

and

$$\begin{aligned}\lambda_1^{-1}(\eta + \sigma^2)^{-1} &= (\lambda_1\eta + \lambda_1\sigma^2)^{-1} \\ &\leq (\lambda_1\eta + \sigma^2)^{-1}\,,\end{aligned}$$

we have

$$g_1(\sigma^2, \eta) \leq g_2(\sigma^2, \eta)\,,$$

where we defined

$$g_1(\sigma^2, \eta) := \lambda_1^{-1}(\eta + \sigma^2)^{-1}||\mathbf{y}||_2^2 + n\log(\lambda_0) + n\log(\eta + \sigma^2)\,.$$

Therefore, we have

$$\min_{\sigma^2,\eta} g_1(\sigma^2, \eta) \leq \min_{\sigma^2,\eta} g_2(\sigma^2, \eta) \leq \min_{\sigma^2,\eta,\mathbf{l}} \mathrm{NLL}(\sigma^2, \eta, \mathbf{l})\,. \tag{3}$$

Next, we will show that, if $||\mathbf{y}||_2^2 \to \infty$, then

$$\min_{\sigma^2,\eta} g_1(\sigma^2, \eta) \to \infty\,.$$

First, note that g_1 depends only on the sum $\eta + \sigma^2$, rather than the individual values. Therefore, we can re-parameterize g_1 as follows

$$g_{1*}(z) := \lambda_1^{-1} z||\mathbf{y}||_2^2 + n\log(\lambda_0) - n\log z\,,$$

where $z := (\eta + \sigma^2)^{-1}$, and we have

$$\min_{z} g_{1*}(z) = \min_{\sigma^2,\eta} g_1(\sigma^2, \eta)\,.$$

Since g_{1*} is a convex function, the minimum value of g_{1*} is attained for $\hat{z}$ with

$$\frac{\partial g_{1*}}{\partial z}(\hat{z}) = \frac{||\mathbf{y}||_2^2}{\lambda_1} - \frac{n}{\hat{z}} = 0\,,$$

and therefore

$$\hat{z} = n\frac{\lambda_1}{||\mathbf{y}||_2^2}\,,$$

and

$$\min_{z} g_{1*}(z) = n + n\log(\lambda_0) - n\log(\lambda_1 n) + n\log(||\mathbf{y}||_2^2)\,.$$

Therefore, if $||\mathbf{y}||_2^2 \to \infty$,

$$\min_{z} g_{1*}(z) \to \infty\,,$$

and as a consequence, from Inequalities (3), we have

$$\min_{\sigma^2,\eta,\mathbf{l}} \mathrm{NLL}(\sigma^2, \eta, \mathbf{l}) \to \infty\,.$$

Therefore, as long as one or more observations belonging to V are selected, we must have that $\min_{\sigma^2,\eta,\mathbf{l}} \mathrm{NLL}(\sigma^2, \eta, \mathbf{l}) \to \infty$. Since $\mathrm{NLL}(\sigma^2, \eta, \mathbf{l})$ is bounded from above for observations belonging to U, the trimmed marginal likelihood GP will select only observations from U.

A.3 ASYMPTOTIC BIAS CORRECTION FOR σ^2

Here, we explain the asymptotic correction for estimating the noise variance for Algorithm 2 in the main paper.

The derivation presented here, generalizes the derivation for the correction of the median linear regression Rousseeuw [1984]. Let Q_f denote the quantile function for distribution f, and by $Q_{\{r_i^2\}_{i=1}^n}$ the empirical quantile function of observed squared residuals r_i^2. We define $Q_{\{r_i^2\}_{i=1}^n}(p) = r^2_{(\lfloor pn \rfloor)}$, where $r^2_{(1)} \leq r^2_{(2)} \ldots \leq r^2_{(n)}$. Let ν be the user-set maximum outlier-ratio, i.e. $1 - \nu = \frac{m}{n}$. Furthermore, note that each r_i^2 is distributed according to $\sigma^2\chi^2(1)$, where $\chi^2(1)$ is the χ^2 distribution with 1 degree of freedom. For $n \to \infty$, we have, see e.g. [Walker, 1968],

$$Q_{\{r_i^2\}_{i=1}^n}(1-\nu) \xrightarrow{p} Q_{\sigma^2\chi^2(1)}(1-\nu)\,.$$

Therefore, for sufficiently large n, we have that

$$\begin{aligned}Q_{\{r_i^2\}_{i=1}^n}(1-\nu) &\approx Q_{\sigma^2\chi^2(1)}(1-\nu)\\ &= \sigma^2 Q_{\chi^2(1)}(1-\nu)\,.\end{aligned}$$

The last line follows from properties of the quantile function (see for example Lemma 1 in this supplement material). Therefore, we set

$$\sigma^2 = \frac{r^2_{(\lfloor(1-\nu)n\rfloor)}}{Q_{\chi^2(1)}(1-\nu)}\,.$$

Lemma 1. *Let Q_X be the quantile function of a real valued random variable X, and define $Y := \alpha X$, where $\alpha > 0$. Then the following holds*

$$Q_Y = \alpha Q_X\,.$$

Proof. First note that

$$\begin{aligned}P(Y \leq y) &= P(X\alpha \leq y)\\ &= P(X \leq \frac{y}{\alpha})\,.\end{aligned}$$

For any $u \in]0,1[$, we have

$$\begin{aligned}Q_Y(u) &= \inf\{y \in \mathbb{R} \,|\, u \leq P(Y \leq y)\}\\ &= \inf\{y \in \mathbb{R} \,|\, u \leq P(X \leq \frac{y}{\alpha})\}\\ &= \alpha \inf\{\frac{y}{\alpha} \in \mathbb{R} \,|\, u \leq P(X \leq \frac{y}{\alpha})\}\\ &= \alpha \inf\{x \in \mathbb{R} \,|\, u \leq P(X \leq x)\}\\ &= \alpha Q_X(u)\,.\end{aligned}$$

□

B DETAILS OF GREEDY METHOD

The function starts with the index set of all data points $S := \{1, 2, \ldots, n\}$, and then removes the data point i_* which leads to the largest marginal likelihood, i.e.

$$i_* := \underset{i \in S}{\arg\max} \left(\log p(\mathbf{y}_{S\setminus\{i\}} | X_{S\setminus\{i\}}, \boldsymbol{\theta}) \right). \tag{4}$$

This is repeated until $|S| = \lceil (1-\nu)n \rceil$. Naively solving the optimization in Equation (4) is in $O(n^4)$, since we need to repeat n-times the calculation of the determinant and inverse of $K_{S\setminus\{i\}}$, where $K_{S\setminus\{i\}}$ denotes the covariance matrix (plus $\sigma^2 I$) of the data points in $S \setminus \{i\}$. However, using the block matrix inversion lemma (together with the Woodbury formula) and the cofactor representation of the determinant, we can solve it in $O(n^3)$ as follows. Without loss of generality assume that sample i corresponds to the last row and column of K_S and write

$$K_S =: \begin{pmatrix} A & \mathbf{b} \\ \mathbf{b}^T & c \end{pmatrix}, \text{and} \quad K_S^{-1} =: \begin{pmatrix} U & \mathbf{v} \\ \mathbf{v}^T & w \end{pmatrix}.$$

Using the block matrix inversion lemma, we have

$$\begin{aligned}U &= A^{-1} + A^{-1}\mathbf{b}(-\mathbf{v}^T)\\ &= A^{-1}(I - \mathbf{b}\mathbf{v}^T)\,,\end{aligned}$$

and therefore

$$\begin{aligned} A^{-1} &= U(I - \mathbf{b}\mathbf{v}^T)^{-1} \\ &= U(I + \mathbf{b}\mathbf{v}^T \frac{1}{1 - \mathbf{v}^T\mathbf{b}})\,, \end{aligned}$$

where in the last line we used the Woodbury formula. Since $A = K_{S\setminus\{i\}}$, this allows for an efficient calculation of $K_{S\setminus\{i\}}^{-1}$. Finally, the determinant $|K_{S\setminus\{i\}}|$ can also be efficiently calculated as follows. Denote the the cofactor matrix of K_S as C, therefore we have $C_{nn} = |A|$. Using the cofactor representation of the inverse, we have

$$K_S^{-1} = \frac{1}{|K_S|} C\,,$$

and therefore

$$\begin{aligned} |A| &= C_{nn} \\ &= |K_S|(K_S^{-1})_{nn}\,. \end{aligned}$$

C COMMENT ON BIAS MODEL FROM PREVIOUS WORKS

The method in [Park et al., 2021] ("Constant Bias Model", Section 3.1) introduces a bias vector $\boldsymbol{\delta} \in \mathbb{R}^n$, where n is the number of samples. If $\delta_i \neq 0$, then sample i is considered an outlier. Furthermore, introducing a Laplace prior on each δ_i, with common scale λ, they propose to jointly estimate $\boldsymbol{\delta}$ and λ as follows:

$$\hat{\boldsymbol{\delta}}, \hat{\lambda} = \arg\min_{\boldsymbol{\delta}, \lambda} \frac{1}{2}(\mathbf{y} - \boldsymbol{\delta})^T A^{-1}(\mathbf{y} - \boldsymbol{\delta}) + \lambda ||\boldsymbol{\delta}||_1 - \log \lambda\,,$$

for some positive definite matrix A, and responses $\mathbf{y} \in \mathbb{R}^n$.[2] They suggest to alternate between the optimization of $\boldsymbol{\delta}$ and λ. However, even only one outlier can lead to a $\hat{\boldsymbol{\delta}}$ which has no zero entry, that is all samples are treated as outliers. To see this, first consider the optimization of $\boldsymbol{\delta}$, leaving λ fixed. Assume that sample i_* is an outlier with $y_{i_*} \to \infty$, then we have $|\delta_{i_*}| \to \infty$. (On the other hand, if $|\delta_{i_*}|$ were bounded, then y_{i_*} would have an arbitrarily large influence on the marginal likelihood.) Next, consider the optimization of λ, leaving $\boldsymbol{\delta}$ fixed: the problem is convex with the unique minimum at

$$\hat{\lambda} = \frac{1}{||\boldsymbol{\delta}||_1}\,.$$

Note that $\frac{1}{||\boldsymbol{\delta}||_1} < \frac{1}{|\delta_{i_*}|}$. Since $|\delta_{i_*}| \to \infty$, we have that $\hat{\lambda} \to 0$. However, if $\hat{\lambda}$ is close to 0, the penalty $\lambda ||\boldsymbol{\delta}||_1$ will in effect be switched off, leading to $\hat{\boldsymbol{\delta}} = \mathbf{y}$.

D ADDITIONAL DETAILS AND EXPERIMENTS

For all methods, we initialize all hyper-parameters $\boldsymbol{\theta}$ to $\log 2$, except the variance σ^2 which is initialized to 10. For all data, we standardize the response and covariates using the median and and the interquartile range (IQR). For all experiments, we used an Nvidia DGX-2. For the real datasets, for evaluating the predictive performance of all methods, we randomly split the data into training (90%) and test data (10%).

D.1 ADDITIONAL RESULTS

References

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.

[2] The term $\lambda||\boldsymbol{\delta}||_1 - \log \lambda$ is supposed to correspond to a Laplace prior on each component of δ_i. However, note that the resulting penalty on λ, should be $-n \log \lambda$ rather than $-\log \lambda$.

Table 1: Estimated upper bound on outlier ratio ν. Except "no extra outliers", the true ratio of added outliers is 0.1.

	no extra outliers	**uniform**	**focused**	**asym**
bow	0.02 (0.01)	0.08 (0.0)	0.09 (0.02)	0.07 (0.0)
F100	0.03 (0.01)	0.07 (0.01)	0.08 (0.03)	0.08 (0.01)
F400	0.02 (0.0)	0.07 (0.0)	0.1 (0.0)	0.07 (0.0)
body	0.02 (0.0)	0.06 (0.01)	0.06 (0.02)	0.07 (0.01)
house	0.02 (0.0)	0.06 (0.0)	0.06 (0.02)	0.06 (0.0)
spacega	0.03 (0.0)	0.07 (0.0)	0.08 (0.0)	0.07 (0.0)

Table 2: Runtime in minutes of each GP regression method.

	GP	**γ-GP**	**t-GP**	**ν-GP**
no extra added outliers				
bow	**0.06** (0.0)	0.1 (0.0)	0.1 (0.0)	5.93 (0.76)
F100	**0.09** (0.01)	0.13 (0.0)	0.17 (0.0)	4.33 (3.42)
F400	**0.1** (0.01)	0.25 (0.02)	0.31 (0.03)	2.88 (0.6)
body	**0.1** (0.0)	0.27 (0.0)	0.23 (0.0)	67.3 (0.0)
house	**0.12** (0.0)	0.25 (0.0)	0.36 (0.0)	17.85 (0.0)
spacega	**1.02** (0.0)	8.88 (0.0)	8.79 (0.0)	9.05 (0.0)
uniform outliers				
bow	**0.06** (0.0)	0.1 (0.0)	0.1 (0.0)	3.44 (0.44)
F100	**0.09** (0.0)	0.13 (0.01)	0.17 (0.0)	2.53 (1.32)
F400	**0.11** (0.01)	0.24 (0.01)	0.15 (0.01)	3.04 (1.23)
body	0.77 (0.48)	0.25 (0.01)	**0.22** (0.0)	29.44 (15.93)
house	0.41 (0.38)	**0.24** (0.02)	**0.24** (0.03)	25.76 (29.33)
spacega	**0.76** (0.02)	8.8 (0.06)	8.78 (0.07)	9.07 (0.17)
focused outliers				
bow	**0.06** (0.0)	0.1 (0.0)	0.1 (0.0)	3.51 (0.53)
F100	**0.09** (0.01)	0.13 (0.0)	0.17 (0.01)	3.24 (2.37)
F400	**0.1** (0.0)	0.23 (0.0)	0.12 (0.02)	4.71 (1.13)
body	**0.1** (0.0)	0.24 (0.01)	0.22 (0.0)	55.5 (43.44)
house	**0.11** (0.0)	0.23 (0.01)	0.28 (0.01)	20.09 (4.42)
spacega	**0.84** (0.01)	8.74 (0.11)	8.67 (0.09)	23.81 (3.73)
asymmetric outliers				
bow	**0.06** (0.0)	0.1 (0.0)	0.1 (0.01)	3.33 (0.36)
F100	**0.09** (0.0)	0.13 (0.01)	0.17 (0.0)	3.66 (3.32)
F400	**0.12** (0.02)	0.23 (0.01)	0.15 (0.02)	2.68 (0.46)
body	0.46 (0.42)	0.24 (0.03)	**0.22** (0.0)	26.23 (14.72)
house	0.3 (0.38)	0.24 (0.01)	**0.23** (0.01)	9.58 (4.56)
spacega	**0.76** (0.02)	8.8 (0.06)	8.78 (0.08)	8.92 (0.23)

Table 3: Runtime in minutes of each optimization method.

	PGD	Greedy (batch)	Greedy (1-by-1)
no extra added outliers			
bow	**0.2** (0.02)	10.37 (7.07)	169.51 (32.26)
F100	**0.14** (0.12)	8.86 (7.98)	5.01 (3.68)
F400	**0.12** (0.05)	10.89 (9.67)	173.58 (52.01)
body	**1.49** (0.0)	3.4 (0.0)	27.17 (0.0)
house	**0.27** (0.0)	7.29 (0.0)	76.35 (0.0)
spacega	**0.82** (0.0)	23.8 (0.0)	-
uniform outliers			
bow	**0.14** (0.04)	2.37 (0.29)	160.39 (3.15)
F100	**0.13** (0.15)	1.74 (1.85)	7.76 (5.66)
F400	**0.15** (0.06)	2.59 (1.44)	42.53 (4.65)
body	**0.79** (0.75)	5.17 (3.97)	65.61 (59.16)
house	**0.21** (0.26)	2.82 (2.64)	150.36 (107.69)
spacega	**0.6** (0.15)	8.52 (0.11)	-
focused outliers			
bow	**0.17** (0.01)	3.49 (0.78)	170.7 (26.81)
F100	**0.14** (0.18)	1.37 (1.06)	8.13 (4.43)
F400	**0.13** (0.0)	2.94 (0.62)	139.74 (19.06)
body	**0.21** (0.24)	2.03 (0.82)	33.37 (37.42)
house	**0.71** (1.19)	6.12 (7.69)	227.69 (209.95)
spacega	**0.9** (0.07)	9.09 (1.44)	-
asymmetric outliers			
bow	**0.09** (0.0)	2.2 (0.07)	48.24 (1.08)
F100	**0.13** (0.15)	2.61 (3.26)	5.23 (3.56)
F400	**0.13** (0.0)	2.13 (1.34)	42.8 (5.34)
body	**0.41** (0.48)	3.18 (4.07)	42.09 (37.34)
house	**0.15** (0.1)	1.3 (1.0)	73.9 (69.2)
spacega	**0.54** (0.01)	8.47 (0.35)	-

Table 4: Marginal likelihood of solution found by different optimization methods.

	PGD	**Greedy (batch)**	**Greedy (1-by-1)**
no extra added outliers			
bow	**1.76** (0.09)	1.75 (0.09)	**1.76** (0.08)
F100	**0.07** (0.12)	-0.06 (0.24)	-0.0 (0.46)
F400	0.34 (0.2)	0.36 (0.23)	**0.42** (0.24)
body	**3.35** (0.0)	3.11 (0.0)	3.23 (0.0)
house	0.11 (0.0)	0.09 (0.0)	**0.18** (0.0)
spacega	-0.31 (0.0)	**0.38** (0.0)	-
uniform outliers			
bow	**1.7** (0.07)	1.54 (0.08)	**1.7** (0.07)
F100	0.01 (0.18)	-0.1 (0.14)	**0.1** (0.17)
F400	0.19 (0.12)	0.07 (0.19)	**0.2** (0.12)
body	**-1.34** (2.33)	-1.5 (2.02)	**-1.34** (2.3)
house	-1.99 (1.13)	-2.0 (1.11)	**-1.96** (1.16)
spacega	-0.26 (0.03)	**0.05** (0.07)	-
focused outliers			
bow	**1.8** (0.05)	1.57 (0.05)	**1.8** (0.05)
F100	0.13 (0.13)	-0.08 (0.25)	**0.22** (0.13)
F400	0.15 (0.04)	-0.0 (0.05)	**0.22** (0.16)
body	0.72 (1.19)	0.46 (0.91)	**0.74** (1.25)
house	0.27 (0.18)	0.15 (0.26)	**0.32** (0.25)
spacega	-0.26 (0.01)	**-0.02** (0.14)	-
asymmetric outliers			
bow	**1.67** (0.1)	1.49 (0.11)	**1.67** (0.1)
F100	**0.15** (0.13)	-0.13 (0.21)	0.14 (0.32)
F400	0.17 (0.07)	0.03 (0.14)	**0.23** (0.13)
body	-1.17 (2.25)	-1.56 (1.5)	**-1.14** (2.27)
house	**-1.23** (0.96)	-1.29 (0.92)	**-1.23** (0.96)
spacega	-0.25 (0.02)	**-0.07** (0.09)	-

Table 5: Outlier ranking performance (R-precision) of different optimization methods.

	PGD	**Greedy (batch)**	**Greedy (1-by-1)**
uniform outliers			
bow	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
F100	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
F400	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
body	**0.87** (0.06)	0.86 (0.06)	0.86 (0.06)
house	**0.86** (0.06)	0.85 (0.06)	**0.86** (0.05)
spacega	0.98 (0.0)	**0.99** (0.01)	-
focused outliers			
bow	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
F100	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
F400	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
body	**1.0** (0.01)	0.95 (0.11)	0.98 (0.05)
house	**0.91** (0.16)	0.55 (0.24)	0.71 (0.32)
spacega	**0.97** (0.0)	0.31 (0.3)	-
asymmetric outliers			
bow	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
F100	**1.0** (0.0)	0.99 (0.03)	**1.0** (0.0)
F400	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
body	**0.86** (0.06)	**0.86** (0.06)	**0.86** (0.06)
house	**0.85** (0.05)	**0.85** (0.05)	**0.85** (0.05)
spacega	0.98 (0.0)	**0.99** (0.0)	-

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

Chiwoo Park, David J Borth, Nicholas S Wilson, Chad N Hunter, and Fritz J Friedersdorf. Robust gaussian process regression with a bias model. *Pattern Recognition*, page 108444, 2021.

Peter J Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

AM Walker. A note on the asymptotic distribution of sample quantiles. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(3):570–575, 1968.

Table 6: Root mean squared error (RMSE) on test data of different optimization methods.

	PGD	**Greedy (batch)**	**Greedy (1-by-1)**
no extra added outliers			
bow	**0.06** (0.0)	**0.06** (0.0)	**0.06** (0.0)
F100	**0.32** (0.05)	0.34 (0.08)	0.42 (0.19)
F400	0.25 (0.05)	**0.23** (0.06)	0.24 (0.05)
body	0.08 (0.1)	**0.05** (0.08)	0.08 (0.09)
house	0.55 (0.12)	**0.46** (0.11)	0.54 (0.13)
spacega	0.49 (0.03)	**0.37** (0.02)	-
uniform outliers			
bow	**0.05** (0.0)	0.06 (0.0)	**0.05** (0.0)
F100	0.31 (0.06)	0.31 (0.07)	**0.29** (0.07)
F400	0.25 (0.03)	**0.23** (0.05)	0.24 (0.03)
body	**0.05** (0.08)	**0.05** (0.07)	**0.05** (0.07)
house	0.4 (0.14)	**0.37** (0.12)	0.4 (0.13)
spacega	0.4 (0.02)	**0.36** (0.01)	-
focused outliers			
bow	**0.05** (0.0)	**0.05** (0.0)	**0.05** (0.0)
F100	0.26 (0.06)	0.3 (0.14)	**0.25** (0.05)
F400	0.25 (0.01)	0.25 (0.01)	**0.24** (0.03)
body	**0.07** (0.08)	0.1 (0.09)	0.08 (0.08)
house	0.4 (0.07)	**0.34** (0.06)	0.39 (0.09)
spacega	**0.41** (0.06)	0.43 (0.04)	-
asymmetric outliers			
bow	**0.06** (0.0)	**0.06** (0.0)	**0.06** (0.0)
F100	**0.26** (0.05)	0.33 (0.09)	0.3 (0.12)
F400	0.25 (0.02)	**0.24** (0.04)	**0.24** (0.03)
body	**0.12** (0.11)	0.15 (0.11)	**0.12** (0.12)
house	0.35 (0.13)	**0.33** (0.09)	0.34 (0.12)
spacega	0.4 (0.02)	**0.37** (0.02)	-