



ISE

Industrial and
Systems Engineering

Nonnegativity-Enforced Gaussian Process Regression

ANDREW PENSONEAULT¹, XIU YANG², AND XUEYU ZHU¹

¹Department of Mathematics, University of Iowa, Iowa, IA 52246, USA

²Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA

ISE Technical Report 20T-033



LEHIGH
UNIVERSITY.



Contents lists available at ScienceDirect

Theoretical & Applied Mechanics Letters

journal homepage: www.elsevier.com/locate/taml

Letter

Nonnegativity-enforced Gaussian process regression

Andrew Pensoneault^a, Xiu Yang^{b,*}, Xueyu Zhu^{a,*}^a Department of Mathematics, University of Iowa, Iowa, IA 52246, USA^b Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015, USA

HIGHLIGHTS

- This method imposes non-negativity in Gaussian process regression in a probabilistic way.
- This method significantly narrows the confident interval in the prediction.

ARTICLE INFO

Article history:

Received 4 February 2020

Received in revised form 20 March 2020

Accepted 27 March 2020

Available online 7 May 2020

Keywords:

Gaussian process regression

Constrained optimization

ABSTRACT

Gaussian process (GP) regression is a flexible non-parametric approach to approximate complex models. In many cases, these models correspond to processes with bounded physical properties. Standard GP regression typically results in a proxy model which is unbounded for all temporal or spacial points, and thus leaves the possibility of taking on infeasible values. We propose an approach to enforce the physical constraints in a probabilistic way under the GP regression framework. In addition, this new approach reduces the variance in the resulting GP model.

©2020 The Authors. Published by Elsevier Ltd on behalf of The Chinese Society of Theoretical and Applied Mechanics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

In many applications, evaluating a computational model can require significant computational resources and time. One approach to address this problem is to build a surrogate model with statistical emulators such as Gaussian processes (GP) regression [1]. We aim to design surrogate models that have low approximation error, and satisfy meaningful bounds on some physical properties. However, no such information is encoded in the standard GP regression method. Therefore, it can produce infeasible predictions.

Incorporating physical information in GP has been explored in many works of literature. For example, it is demonstrated in Ref. [2] that the mean prediction of a GP model satisfies a set of linear equality constraints provided the training data satisfy these constraints. A similar result holds for quadratic equality constraints under a transformation of the parameterization. Alternatively, linear equality constraints can be enforced by modeling the process as a transformation of an underlying function and imposing the constraints on that transformation [3].

* Corresponding author.

E-mail address: xiy518@lehigh.edu (X. Yang), xueyu-zhu@uiowa.edu (X. Zhu).

Moreover, physical information in the form of differential operators can be incorporated in GP models [4–7].

Incorporating inequality constraints in a GP is more difficult, as the underlying process conditional on the constraints is no longer a GP [8]. To address this problem, several different approaches have been explored. The approach in Ref. [9] enforces inequality constraints at several locations and draws approximate samples from the predictive distribution with a data augmentation approach. Linear inequality functional (such as monotonicity) are enforced via virtual observations at several location within [10–14]. In Ref. [11], it is shown that when linear inequality constraints are applied to a finite set of points in the domain, the process conditional on the constraints is a compound GP with a truncated Gaussian mean. In Ref. [8, 15, 16], linear inequality constraints are enforced on the entire domain instead of a finite set of points by making a finite-dimensional approximation of the GP and enforcing the constraints through the choice of the associated approximation coefficients.

In this work, we focus on enforcing non-negativity in the GP model. This is a requirement for many physical properties, e.g., elastic modulus, viscosity, density, and temperature. We propose to impose this inequality constraint with high probability

via selecting a set of constraint points in the domain and imposing the non-negativity on the posterior GP at these points. In addition to enforcing non-negativity, this approach improves accuracy and reduces uncertainty in the resulting GP model.

The paper is organized as follows. We review the standard GP regression framework first, then present our novel approach to enforce non-negativity in GP regression, and provide numerical examples at last.

We introduce the framework for GP regression based on the descriptions in Ref. [17]. Assume we have $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(N)})^T$ as the values of the target function, where $y^{(i)} \in \mathbb{R}$ are observations at locations $\mathbf{X} = \{x^{(i)}\}_{i=1}^N$ where $x^{(i)}$ are d -dimensional vectors in the domain $D \in \mathbb{R}^d$. We aim to use a GP $Y(\cdot, \cdot) : D \times \Omega \rightarrow \mathbb{R}$ to approximate the underlying target function. Typically, $Y(x)$ is denoted as

$$Y(x) \sim \mathcal{GP}[\mu(x), \mathcal{K}(x, x')], \quad (1)$$

where $\mu(\cdot) : D \rightarrow \mathbb{R}$ and $\mathcal{K}(\cdot, \cdot) : D \times D \rightarrow \mathbb{R}$ are the associated mean function and covariance function, i.e

$$\mu(x) = \mathbb{E}[Y(x)], \quad (2)$$

$$\mathcal{K}(x, x') = \mathbb{E}[Y(x) - \mu(x)][Y(x') - \mu(x')]. \quad (3)$$

A widely used kernel is the standard squared exponential covariance kernel with an additive independent identically distributed Gaussian noise term ϵ with variance σ_n^2 :

$$\mathcal{K}(x, x') = \sigma^2 \exp\left(\frac{-\|x - x'\|_2^2}{2l^2}\right) + \sigma_n^2 \delta_{x, x'}, \quad (4)$$

where $\delta_{x, x'}$ is a Kronecker delta function, l is the length-scale, and σ^2 is the signal variance. In general, by assuming zero mean function $\mu(x) \equiv 0$, we use $\theta = (\sigma, l, \sigma_n)$ to denote the hyperparameters, and they are determined based on the training data.

Define $K = [\mathcal{K}(x^{(i)}, x^{(j)})]_{ij}$, $\boldsymbol{\mu} = [\mu(x^{(1)}), \mu(x^{(2)}), \dots, \mu(x^{(N)})]^T$, and $k(x') = [\mathcal{K}(x^{(i)}, x')]_i - \sigma_n^2 \delta_{x^{(i)}, x'}$. The posterior predictive distribution of the output \mathbf{y}^* given the training set is $Y(x^*)|x^*, \mathbf{y}, \mathbf{X} \sim \mathcal{N}[\mathbf{y}^*(x^*), s^2(x^*)]$, where

$$\mathbf{y}^*(x^*) = \boldsymbol{\mu}(x^*) + k(x^*)^T K^{-1}(\mathbf{y} - \boldsymbol{\mu}), \quad (5)$$

$$s^2(x^*) = \sigma^2(x^*) - k(x^*)^T K^{-1}k(x^*). \quad (6)$$

One approach to determine the hyperparameters θ is to minimize the negative marginal log-likelihood [17]:

$$-\lg[p(\mathbf{y}|\mathbf{X}, \theta)] = \frac{1}{2} [(\mathbf{y} - \boldsymbol{\mu})^T K^{-1}(\mathbf{y} - \boldsymbol{\mu}) + \lg|K| + N \lg(2\pi)]. \quad (7)$$

In particular, we enforce the non-negativity in the quantity of interest. We minimize the negative marginal log-likelihood function in Eq. (7) while requiring that the probability of violating the constraints is small. More specifically, for $0 < \eta \ll 1$, we impose the following constraint:

$$P[(Y(x)|x, \mathbf{y}, \mathbf{X}) < 0] \leq \eta, \quad \text{for all } x \in D. \quad (8)$$

This differs from other methods in the literature, which enforce

the constraint via truncated Gaussian assumption [8], or use a bounded likelihood function and perform inference based on the Laplace approximation and expectation propagation [18]. In contrast, our method retains the Gaussian posterior of standard GP regression, and only requires a slight modification of the existing cost function. As $Y(x)|x, \mathbf{y}, \mathbf{X}$ follows a Gaussian distribution, this constraint can be rewritten in terms of the posterior mean \mathbf{y}^* and posterior standard deviation s :

$$y^*(x) + \Phi^{-1}(\eta)s(x) \geq 0, \quad \text{for all } x \in D, \quad (9)$$

where Φ^{-1} is the inverse cumulative density function (CDF) of a standard Gaussian random variable. In this work, we set $\eta = 2.2\%$ for demonstration purpose, and consequently $\Phi^{-1}(\eta) = -2$, i.e., two standard deviations below the mean is still nonnegative. Therefore, we minimize the negative log-likelihood cost function subject to constraints on the posterior mean and standard deviation:

$$\underset{\theta}{\operatorname{argmin}} \quad -\lg[p(\mathbf{y}|\mathbf{X}, \theta)], \quad (10)$$

$$\text{s.t. } y^*(x) - 2s(x) \geq 0 \quad \text{for all } x \in D. \quad (11)$$

We note that Eq. (11) is a functional constraint and thus can be difficult to enforce. Instead, we enforce Eq. (11) on a set of constraint points $\mathbf{X}_c = \{x_c^{(i)}\}_{i=1}^m$. Of note, these constraint points play similar roles as the aforementioned virtual observations [10–14].

Meanwhile, in practice, a heuristic on the distance of the posterior mean of the GP from the training data is applied to stabilize the optimization algorithm, as such to guarantee that it results in a model that fits measurement data. Subsequently, to obtain the constrained GP, we solve the following constrained minimization problem:

$$\underset{\theta}{\operatorname{argmin}} \quad -\lg[p(\mathbf{y}|\mathbf{X}, \theta)], \quad (12)$$

$$\text{s.t. } 0 \leq y^*(x_c^{(i)}) - 2s(x_c^{(i)}), \quad \text{for all } i = 1, 2, \dots, m, \quad (13)$$

$$0 \leq \epsilon - |y^{(j)} - y^*(x^{(j)})|, \quad \text{for all } j = 1, 2, \dots, n, \quad (14)$$

where $\epsilon > 0$ is chosen to be sufficiently small. In this paper, we set $\epsilon = 0.03$. The last constraint is chosen so that the given solution fits the data sufficiently well.

We remark that compared with unconstrained optimization, constrained optimization is in general more computationally expensive [19]. However, if non-negativity approximation of the target function is crucial for the underlying applications, one may weigh less on the efficiency in order to get more reliable and feasible approximation within the computational budget.

We present numerical examples to illustrate the effectiveness of our method. We measure the relative l_2 error between the posterior mean \mathbf{y}^* and the true value of the target function $f(x)$ over a set of test points $\mathbf{X}_T = \{x_T^{(i)}\}_{i=1}^{N_T}$.

$$E = \sqrt{\frac{\sum_{i=1}^{N_T} [y^*(x_T^{(i)}) - f(x_T^{(i)})]^2}{\sum_{i=1}^{N_T} f(x_T^{(i)})^2}}. \quad (15)$$

For the examples below, we use $N_T = 1000$ equidistant test points

over the domain D . We use the standard squared exponential covariance kernel as well as a zero prior mean function $\mu(x) = 0$. We solve the unconstrained log-likelihood minimization problem in MATLAB using the GPML package [20]. For the constrained optimization, we use the fmincon from the MATLAB Optimization Toolbox based on the built-in interior point algorithm [21].

Remark *If the method results in convergence to an infeasible solution, the optimization is performed again with another random initial guess (with a standard Gaussian noise added to the base initial condition, $\theta_0 = [\lg(l), \lg(\sigma), \lg(\sigma_n)] = (-3, -3, -10)$.*

Example 1

Consider the following function:

$$f(x) = \frac{1}{1 + (10x)^4} + \frac{1}{2} \exp \left[-100 \left(x - \frac{1}{2} \right)^2 \right], \quad x \in [0, 1].$$

For our tests on this example, the training point set is

$$\{x^{(i)}\}_{i=1}^7 = \left\{ \frac{j-1}{5} + \epsilon_j \right\}_{j=1}^6 \cup \left\{ \frac{1}{2} \right\},$$

where $\epsilon_j \sim \mathcal{N}(0, 0.03^2)$ for $j = 2, 3, 4, 5$ and $\epsilon_1 = \epsilon_6 = 0$. We chose $m = 30$ equidistant points over the domain as our constraint points.

Figure 1a shows the posterior mean of the unconstrained GP with 95% confidence interval. It can be seen that on $[0.65, 0.85]$, the posterior mean violates the non-negativity bounds with a large variance. In contrast, the posterior mean of the constrained GP in these regions no longer violates the constraints, as shown in Fig. 1b. Besides, the confidence interval is reduced dramatically after the non-negativity constraint is imposed.

To illustrate the robustness of the algorithm, we repeat the same experiment on 100 different training data sets as in Ref. [4]. Figure 2a illustrates the distribution of the relative l_2 error over the 100 trials. It is clear that incorporating the constraint tends to result in a lower relative error in the posterior mean statistically. Figure 2b compares the percentage of the posterior mean over the test points that violate non-negativity constraint over the 100 trials. There is a large portion of the posterior mean by the un-

constrained GP that violates the non-negativity, while the constrained GP preserves the non-negativity very well.

Example 2

Consider the following function:

$$f(x) = \frac{1}{100} + \frac{5}{8} (2x - 1)^4 [(2x - 1)^2 + 4 \sin(5\pi x)^2], \quad x \in [0, 1].$$

We train our constrained and unconstrained GP models over 14 training points at locations:

$$\{x^{(i)}\}_{i=1}^{14} = \left\{ \frac{j-1}{11} + \epsilon_j \right\}_{j=1}^{12} \cup \{0.075, 0.925\},$$

where $\epsilon_j \sim \mathcal{N}(0, 0.03^2)$ for $j = 2, 3, \dots, 11$ and $\epsilon_1 = \epsilon_{12} = 0$. We choose $m = 31$ equidistant points in the domain as our constraint points.

Figure 3a shows a 95% confidence interval around the posterior mean of the unconstrained GP. Notice that the posterior mean is less than zero near neighborhoods of 0.8. In contrast, the constrained GP doesn't violate the constraints as shown in Fig. 3b. The confidence interval of the posterior mean is also much narrower, which illustrates the advantage of incorporating the constraints.

Again, to show the robustness of the algorithm, we repeat the same experiment on 100 trials. Figure 4a shows the relative l_2 error over 100 trials. The constrained GP has a histogram more heavily weighted towards lower relative error in the posterior mean, compared to the unconstrained GP. Figure 4b shows that the posterior mean of the unconstrained GP violates the non-negativity condition more frequently.

Example 3

The Korteweg–de Vries (KdV) [22] equation can be used to describe the evolution of solitons, which are characterized by the following properties: (1) invariant shape; (2) approaches a constant as $t \rightarrow \infty$; (3) strong interactions with other solitons. We consider the KdV equation in the following form

$$u_t(x, t) - 6u(x, t)u_x(x, t) + u_{xxx}(x, t) = 0.$$

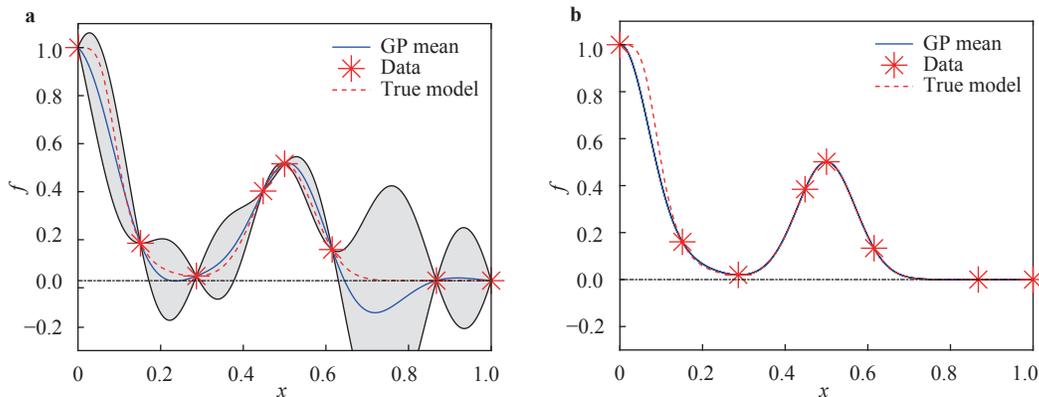


Fig. 1. Posterior mean and the corresponding 95% confidence interval of the GP models in example 1. **a** Unconstrained GP. **b** Constrained GP.

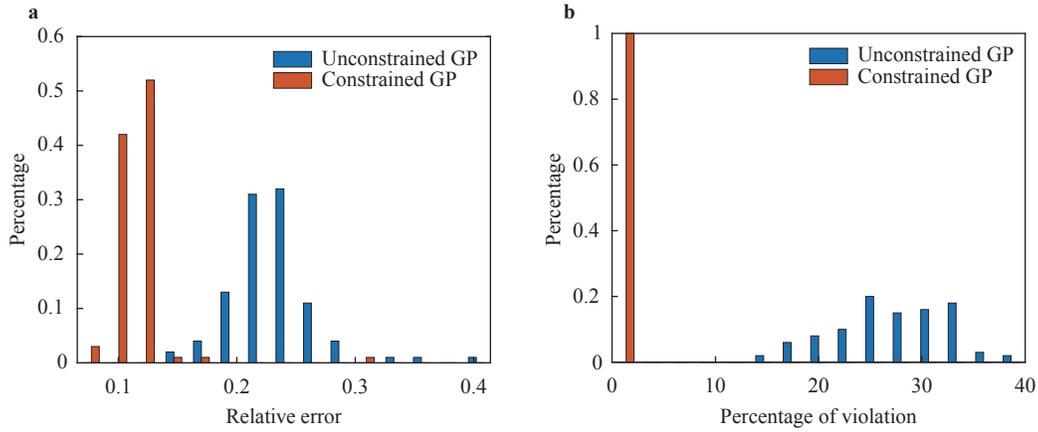


Fig. 2. **a** Normalized histogram associated with the l_2 relative error between the GP mean and the true function over the test set based on 100 different training sets. **b** Normalized histogram associated with the percentage of the posterior mean over test points that violate the non-negativity constraint.

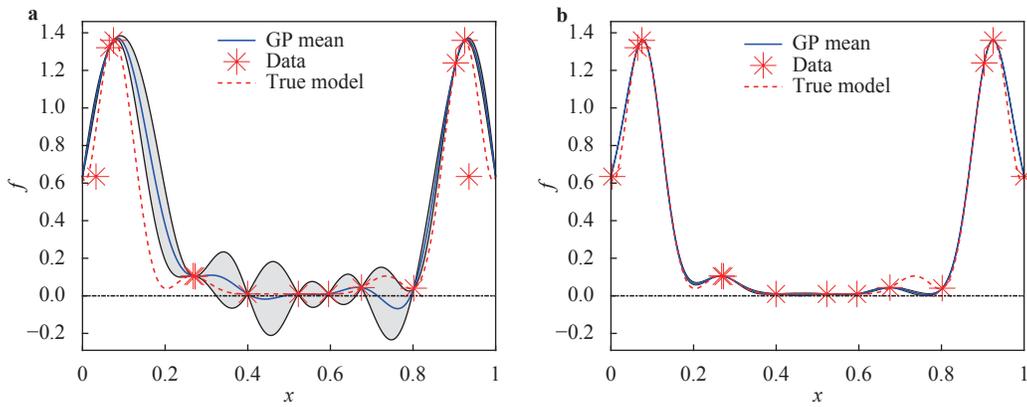


Fig. 3. Posterior mean and the corresponding 95% confidence interval of the GP models in example 2. **a** Unconstrained GP. **b** Constrained GP.

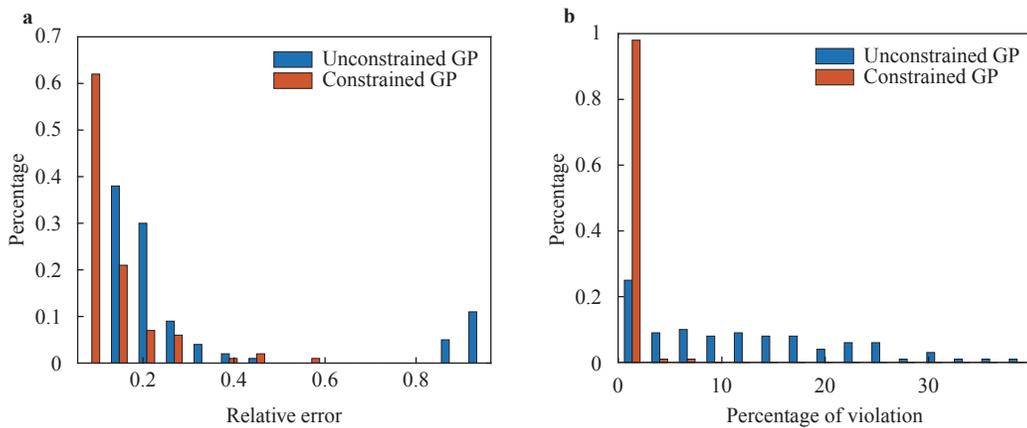


Fig. 4. **a** Normalized histogram associated with the l_2 relative error between the GP mean and the true function over the test set based on 100 different training sets. **b** Normalized histogram associated with the percentage of the posterior mean over test points which violate the non-negativity constraint.

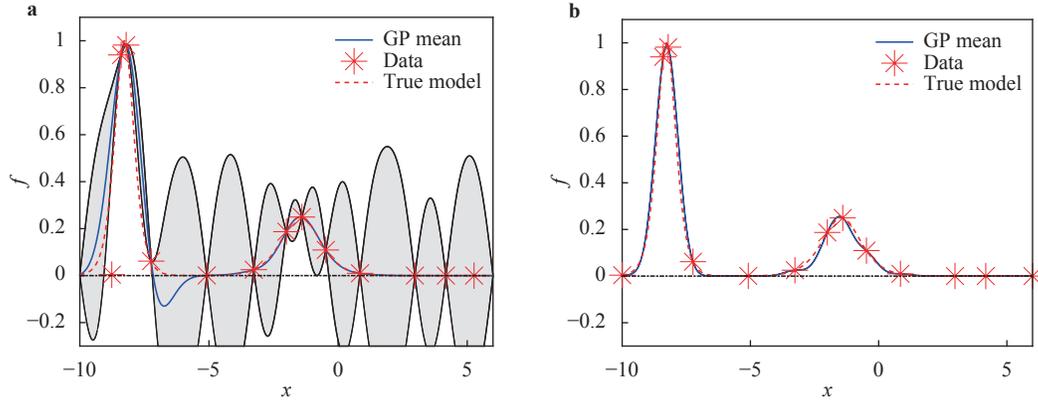


Fig. 5. Posterior mean and the corresponding 95% confidence interval of GP models approximating the two-soliton interacting system at $t = -1$ for a set training data set. **a** Unconstrained GP. **b** Constrained GP.

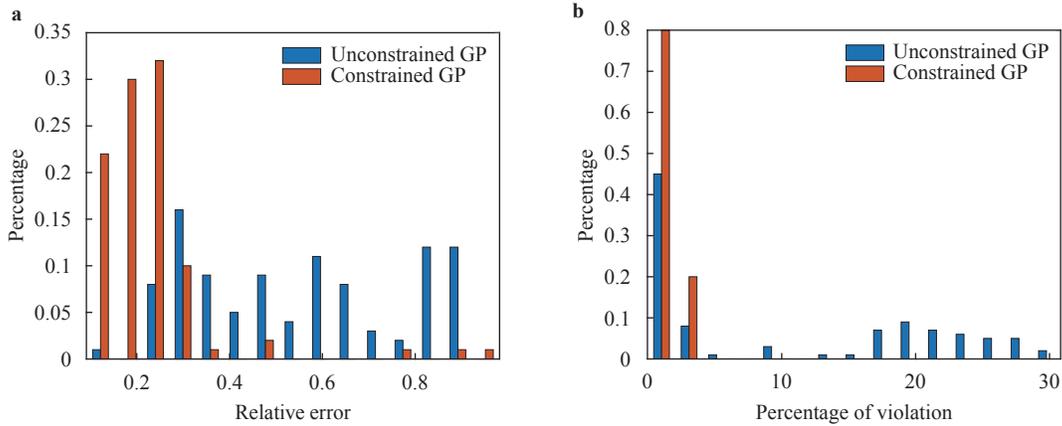


Fig. 6. **a** Normalized histogram associated with the l_2 relative error between the GP mean and the true function over the test set based on 100 different training sets. **b** Normalized histogram associated with the percentage of the posterior mean over test points that violate the non-negativity constraint.

Under several assumptions on the form of u , an analytic solution can be found. For the case of two solitons, a (normalised) solution can be found in Ref. [22]:

$$u(x, t) = \frac{12[3 + 4 \cosh(2x - 8t) + \cosh(4x - 64t)]}{8[3 \cosh(x - 28t) + \cosh(3x - 36t)]^2}. \quad (16)$$

For this equation, $u(x, t) > 0$ for all $x, t \in \mathbb{R}$, we aim to approximate $u(x, -1)$ using GP.

We train our constrained and unconstrained GP model based on 13 training points at locations:

$$\{x^{(i)}\}_{i=1}^{13} = \left\{ -10 + 15 \frac{j-1}{10} + \epsilon_j \right\}_{j=1}^{11} \cup \{-1.4, -8.4\},$$

where $\epsilon_j \sim \mathcal{N}(0, 0.3^2)$ for $j = 2, 3, \dots, 10$ and $\epsilon_1 = \epsilon_{11} = 0$. We choose $m = 40$ equidistant points in the domain as our constraint points.

As can be seen in Fig. 5, the unconstrained GP violates non-negativity around $x = -7$, which is avoided in the constrained GP. More importantly, the confidence interval of the resulting GP is dramatically reduced by imposing non-negativity constraint. In addition, Fig. 6a shows that the relative error is significantly re-

duced when we incorporate the non-negativity information. Of note, in this case, because the majority of the test points are near zero, the relative error is much more sensitive to approximation errors in these regions. Figure 6b illustrates that the constrained GP preserves the non-negativity with very high probability while the unconstrained GP violates the non-negativity much more frequently.

In this paper, we propose a novel method to enforce the non-negativity constraints on the GP in the probabilistic sense. This approach not only reduces the difference between the posterior mean and the ground truth, but significantly lowers the variance, i.e., narrows the confidence interval, in the resulting GP model because the non-negativity information is incorporated. While this paper covers only the non-negativity bound, other inequality constraints can be enforced in a similar manner.

Acknowledgement

X.Y. Zhu's work was supported by Simons Foundation. X. Yang's work was supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Re-

search as part of Physics-Informed Learning Machines for Multiscale and Multiphysics Problems (PhILMs).

References

- [1] J. Sacks, W.J. Welch, T.J. Mitchell, et al., Design and analysis of computer experiments, *Statist. Sci.* 4 (1989) 409–423.
- [2] M. Salzmann, R. Urtasun, Implicitly constrained Gaussian process regression for monocular non-rigid pose estimation, *Adv. Neural. Inf. Process. Syst.* (2010) 2065–2073.
- [3] C. Jidling, N. Wahlström, A. Wills, et al., Linearly constrained Gaussian processes, *Adv. Neural. Inf. Process. Syst.* (2017) 1215–1224.
- [4] M. Schober, D.K. Duvenaud, P. Hennig, Probabilistic ode solvers with Runge-Kutta means, *Adv. Neural. Inf. Process. Syst.* (2014) 739–747.
- [5] M. Raissi, P. Perdikaris, G.E. Karniadakis, Numerical Gaussian processes for time-dependent and nonlinear partial differential equations, *SIAM J. Sci. Comput.* 40 (2018) A172–A198.
- [6] X. Yang, X. Zhu, J. Li, When bifidelity meets cokriging: An efficient physics-informed multifidelity method, *SIAM J. Sci. Comput.* 42 (2020) A220–249.
- [7] X. Yang, G. Tartakovsky, A. Tartakovsky, Physics-informed kriging: A physics-informed Gaussian process regression method for data-model convergence, arXiv: 1809.03461 (2018).
- [8] H. Maatouk, X. Bay, Gaussian process emulators for computer experiments with inequality constraints, *Math. Geol.* 49 (2017) 557–582.
- [9] P. Abrahamsen, F.E. Benth, Kriging with inequality constraints, *Math. Geol.* 33 (2001) 719–744.
- [10] X. Wang, J.O. Berger, Estimating shape constrained functions using Gaussian processes, *SIAM/ASA J. Uncertain. Quantif.* 4 (2016) 1–25.
- [11] C. Agrell, Gaussian processes with linear operator inequality constraints, arXiv: 1901.03134 (2019).
- [12] S. Golchi, D.R. Bingham, H. Chipman, et al., Monotone function estimation for computer experiments, arXiv: 1309.3802 (2013).
- [13] J. Riihimäki, A. Vehtari, Gaussian processes with monotonicity information, in: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (2010) 645–652.
- [14] S. Da Veiga, A. Marrel, Gaussian process modeling with inequality constraints, *Annales de la Faculte des Sciences de Toulouse* 21 (2012) 529–555.
- [15] A.F. López-Lopera, F. Bachoc, N. Durrande, et al., Finite-dimensional Gaussian approximation with linear inequality constraints, *SIAM/ASA J. Uncertain. Quantif.* 6 (2018) 1224–1255.
- [16] H. Maatouk, O. Roustant, Y. Richef, Cross-validation estimations of hyper-parameters of Gaussian processes with inequality constraints, *Procedia. Environ. Sci.* 27 (2015) 38–44.
- [17] C.E. Rasmussen, M. Kuss, Gaussian processes in reinforcement learning, *Adv. Neural. Inf. Process. Syst.* 16 (2004) 751–759.
- [18] B.S. Jensen, J.B. Nielsen, J. Larsen, Bounded gaussian process regression, in: 2013 IEEE International Workshop on Machine Learning for Signal Processing (2013) 1–6.
- [19] M. Box, A new method of constrained optimization and a comparison with other methods, *The Computer Journal* 8 (1965) 42–52.
- [20] C.E. Rasmussen, H. Nickish, Gaussian processes for machine learning (GPML) toolbox, *J. Mach. Learn. Res.* 11 (2010) 3011–3015.
- [21] MATLAB version 9.6.0.1214997 (R2019a). The MathWorks Inc., Natick, Massachusetts (2019).
- [22] N. Schalch, The Korteweg-de Vries Equation, Proseminar: Algebra, Topology and Group Theory in Physics (2018).