Explaining the Uncertain: Stochastic Shapley Values for Gaussian Process Models

Siu Lun Chau[1], Krikamol Muandet[1]*, Dino Sejdinovic*[2]

[1]CISPA Helmholtz Center for Information Security [2]University of Adelaide

TL;DR

What is GP-SHAP?

1. Similar to TreeSHAP for trees, DeepSHAP for deep model, RKHS-SHAP for kernel methods, **GP-SHAP is a model-specific SHAP algorithm for Gaussian process models**.
2. GP-SHAP formulate explanations as stochastic Shapley values to leverage both **predictive** and **estimation uncertainties** to provide **uncertainties around explanations**.

What is predictive explanation and the Shapley prior?

1. Predictive explanation focuses on predicting **feature contributions for unseen data.**
2. The Shapley prior is an **induced prior** on the space of explanation functions for priors $f \sim \mathcal{GP}(0, k)$.

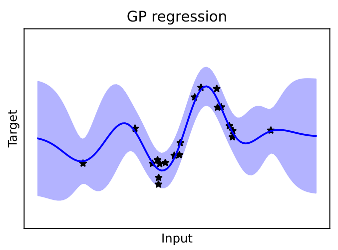
Gaussian process recap

Proposition 1: Gaussian Process regression

Given data $\mathbf{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ and $y_i = f(x_i) + \sigma\mathcal{N}(0,1)$, if we posit a GP prior $\mathcal{GP}(0, k)$ on f, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance kernel, then $f \mid \mathbf{D} \sim \mathcal{GP}(\tilde{m}, \tilde{k})$ where:

$$\tilde{m}(\mathbf{x}') = k(\mathbf{x}', \mathbf{X})(\mathbf{K}_{\mathbf{X}} + \sigma^2 I)^{-1}\mathbf{y}$$

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X})(\mathbf{K}_{\mathbf{X}} + \sigma^2 I)^{-1}k(\mathbf{X}, \mathbf{x}')$$



- The posterior covariance captures uncertainty around predictions.

What is Stochastic Shapley values?

Proposition 2: Stochastic Shapley values

Denote $[d] := 1, \dots, d$ as the player set and $\nu : 2^{[d]} \to \mathcal{L}_2(\mathbb{R})$ a stochastic cooperative game, then the corresponding stochastic Shapley values take the form:

$$\phi_i(\nu) = \sum_{S \subseteq [d] \setminus \{i\}} c_{|S|}\left(\nu(S \cup i) - \nu(S)\right)$$

where $c_{|S|} = \frac{1}{d}\binom{d-1}{|S|}^{-1}$ and $\phi_i(\nu)$ is the i^{th} SSV of game ν.

- Analogous to the standard deterministic SVs except it is written in terms of random variables.

Proposition 3: Variance of Stochastic Shapley values

The Stochastic Shapley value variance is given by $\mathbb{V}[\phi_i(\nu)] =$

$$\sum_{S \subseteq [d] \setminus \{i\}} \sum_{S' \subseteq [d] \setminus \{i\}} c_{|S|}c_{|S'|}\left(\mathbb{C}[\nu(S \cup i), \nu(S' \cup i)] - \mathbb{C}[\nu(S \cup i), \nu(S')] - \mathbb{C}[\nu(S), \nu(S' \cup i)] + \mathbb{C}[\nu(S), \nu(S')]\right),$$

where $\mathbb{C}$ is the covariance function between the stochastic payoffs.

- The mean of SSVs coincide with the SVs of mean games,
- But the variance of SSVs do not coincide with the SVs of variance games.

Explanations of GPs: GP-SHAP

Stochastic cooperative game from GP. Given a posterior GP $f \mid \mathbf{D} \sim \mathcal{GP}(\tilde{m}, \tilde{k})$, a stochastic cooperative game can be formulated as $\nu_f(\mathbf{x}, S) := \mathbb{E}_X[f(X) \mid X_S = \mathbf{x}_S]$, which again is a GP with mean $\tilde{m}_\nu(\mathbf{x}, S) := \mathbb{E}[\tilde{m}(X) \mid X_S = \mathbf{x}_S]$ and covariance $\tilde{k}_\nu((\mathbf{x}, S), (\mathbf{x}', S')) := \mathbb{E}[\tilde{k}(X, X') \mid X_S = \mathbf{x}_S, X'_{S'} = \mathbf{x}'_{S'}]$.

Theorem 1: Stochastic Shapley values for ν_f and how to estimate them.

Let ν_f be an induced stochastic game from the GP $f \sim \mathcal{GP}(\tilde{m}, \tilde{k})$ and denote $\mathbf{v_x} := [\nu_f(\mathbf{x}, S_1), \dots \nu_f(\mathbf{x}, S_{2^d})]^\top$ the vector of stochastic payoffs across all coalitions, then the corresponding stochastic Shapley values $\phi(\nu_f(\mathbf{x}, \cdot))$ follows a d-dimensional multivariate Gaussian distribution,

$$\phi(\nu_f(\mathbf{x}, \cdot)) \sim \mathcal{N}(\mathbf{A}\mathbb{E}[\mathbf{v}_x], \mathbf{A}\mathbb{V}[\mathbf{v}_x]\mathbf{A}^\top) \quad \text{with} \quad \mathbf{A} := (\mathbf{Z}^\top\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{W},$$

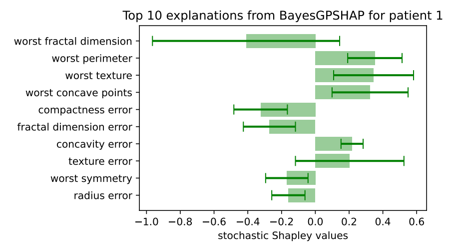
where $\mathbb{E}[\mathbf{v}_x] \in \mathbb{R}^{2^d}$ and $\mathbb{V}[\mathbf{v}_x] \in \mathbb{R}^{2^d \times 2^d}$ are the corresponding mean vector and covariance matrix of the payoffs and $\mathbf{A}$ is the regression operator used in the WLS formulation of Shapley values. The moments for the multivariate stochastic Shapley values can be estimated as,

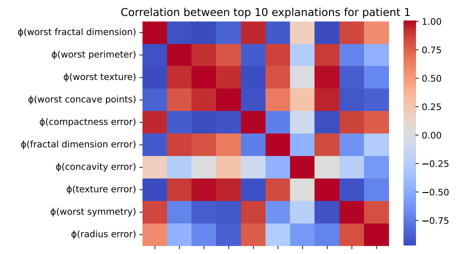
$$\phi\left(\hat{\nu}_f(\mathbf{x}, \cdot)\right) = \mathcal{N}\left(\mathbf{A}\mathbf{B}(\mathbf{x}, [d])^\top\tilde{m}(\mathbf{X}), \mathbf{A}\mathbf{B}(\mathbf{x}, [d])^\top\tilde{\mathbf{K}}_{\mathbf{XX}}\mathbf{B}(\mathbf{x}, [d])\mathbf{A}^\top\right)$$

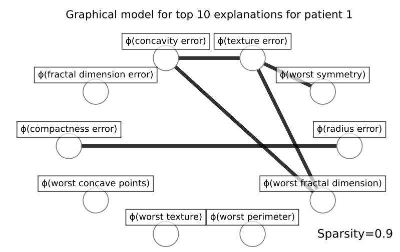
where $\mathbf{B}(\mathbf{x}, [d]) = [\mathbf{b}(\mathbf{x}, [d]_1), \dots, \mathbf{b}(\mathbf{x}, [d]_{2^d})]^\top$, $\mathbf{b}(\mathbf{x}, S) := (\mathbf{K}_{\mathbf{X}_S\mathbf{X}_S} + \lambda I)^{-1}k_S(\mathbf{X}_S, \mathbf{x}_S)$, $\tilde{m}(\mathbf{X}) = [\tilde{m}(\mathbf{x}_1), \dots, \tilde{m}(\mathbf{x}_n)]^\top$. The parameter $\lambda > 0$ is a fixed hyperparameter to stabilise the inversion.

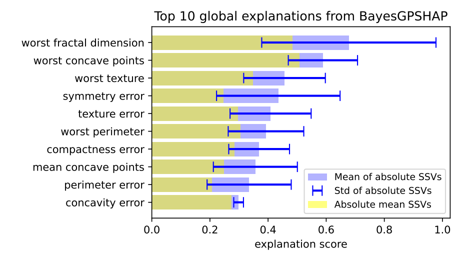
- This captures predictive uncertain from posterior GP. By leveraging the Bayesian weighted regression formulation from Slack et al. 2022, we can in addition integrate the estimation uncertainty to our formulation.

Illustrations. With the fully tractable covariance structure of explanations across both features and observations, we can comprehend the explanations with the following tools.









Explanations as GPs: Predictive explanations with the Shapley prior

Predictive Explanations. Given seen explanations, can we capture the Shapley explanation function $\phi : \mathcal{X} \to \mathbb{R}^d$?

Proposition 4: The Shapley prior

The prior $f \sim \mathcal{GP}(0, k)$ and the game $\nu_f(\mathbf{x}, S) = \mathbb{E}[f(X) \mid X_S = \mathbf{x}_S]$ induce a vector-valued GP prior over the explanation functions $\phi \sim \mathcal{GP}(0, \kappa)$ where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d \times d}$ is the matrix-valued covariance kernel

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathcal{A}(\mathbf{x})^\top\mathcal{A}(\mathbf{x}'), \quad \mathcal{A}(\mathbf{x}) = \Psi(\mathbf{x})\mathbf{A}^\top$$

where $\Psi(\mathbf{x}) = \left[\mathbb{E}[k(\cdot, X) \mid X_{S_1} = x_{S_1}], \dots, \mathbb{E}[k(\cdot, X) \mid X_{S_{2^d}} = x_{S_{2^d}}]\right]$.

Using this vector-valued GP prior, we can treat $\{\mathbf{x}_i, \mathbf{\Phi}_{\mathbf{x}_i}\}_{i=1}^n$ as regression data. Furthermore, the posterior mean of this vector-valued GP corresponds to Shapley values of certain payoffs,

Proposition 5: Posterior mean as Shapley values of certain payoffs

The posterior mean $\tilde{m}_\phi(\mathbf{x}')$ corresponds to Shapley values for the payoff vector $\tilde{\mathbf{v}}_{\mathbf{x}'}$, i.e., $\tilde{m}_\phi(\mathbf{x}') = \mathbf{A}\tilde{\mathbf{v}}_{\mathbf{x}'}$, where $\tilde{\mathbf{v}}_{\mathbf{x}'} = \sum_{i=1}^n \Psi(\mathbf{x}')^\top\Psi(\mathbf{x}_i)\mathbf{A}^\top\alpha_i$ and $\alpha_i \in \mathbb{R}^d$ is the $[i, ..., i + (d-1)]$ subvector of $(\kappa_{\mathbf{XX}} + \sigma_\phi^2 I)^{-1}\operatorname{vec}(\mathbf{\Phi}_{\mathbf{X}})$.

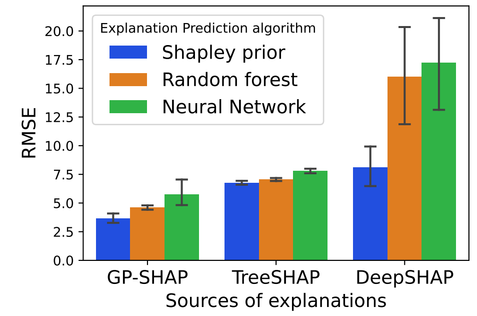


Figure 1: Predicting explanations generated using different explanation algorithms