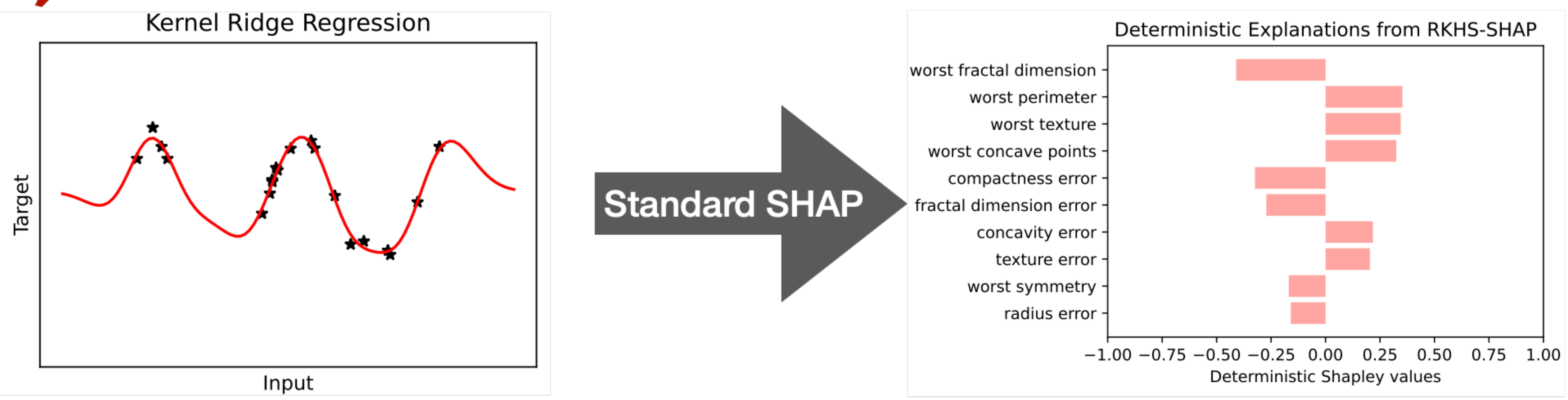
Explaining the Uncertain: Stochastic Shapley Values for Gaussian Process Models

NeurIPS 2023 Spotlight Poster

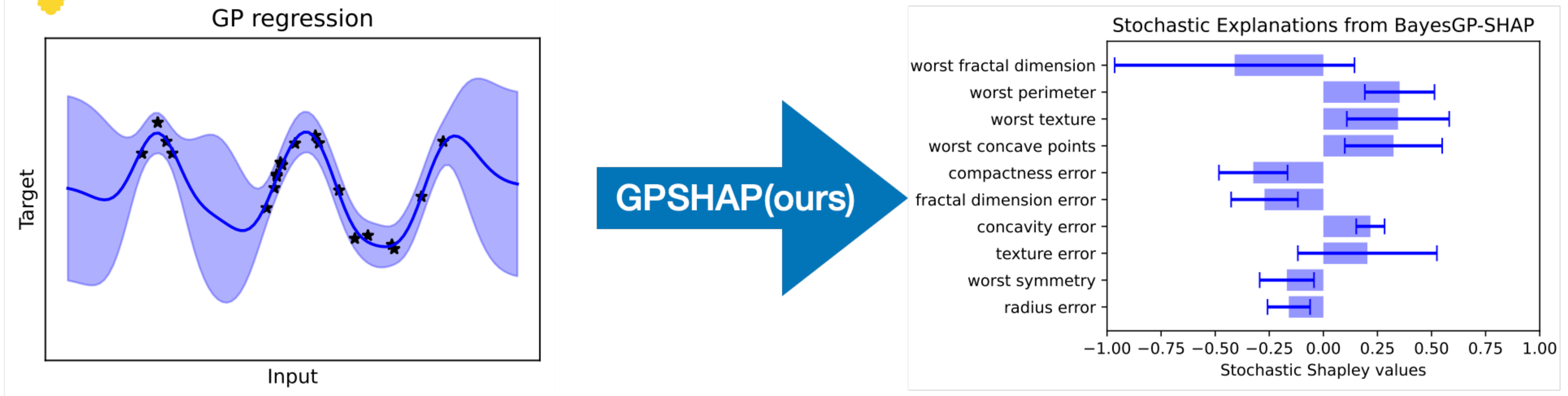
Siu Lun Chau[1], Krikamol Muandet[1]*, Dino Sejdinovic*[2]
[1]CISPA Helmholtz Center for Information Security, Germany [2]University of Adelaide, Australia

MOTIVATION

While deterministic model gives deterministic explanations....

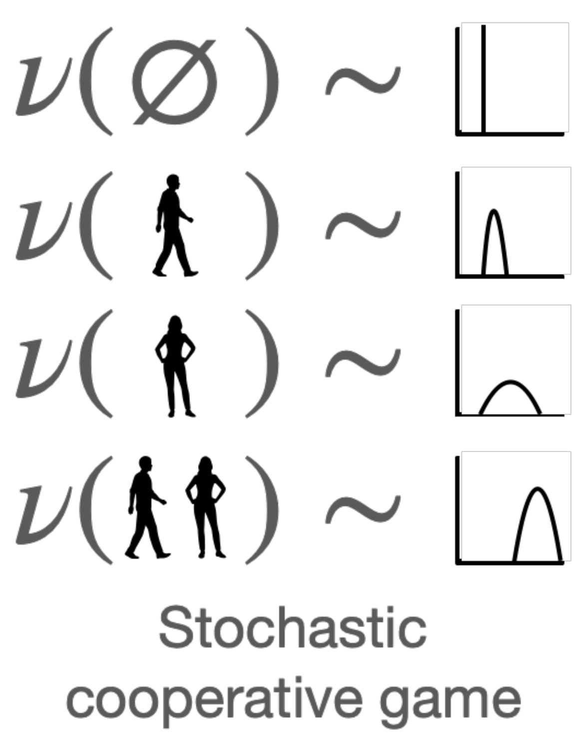


Shouldn't probabilistic models get their stochastic explanations too?



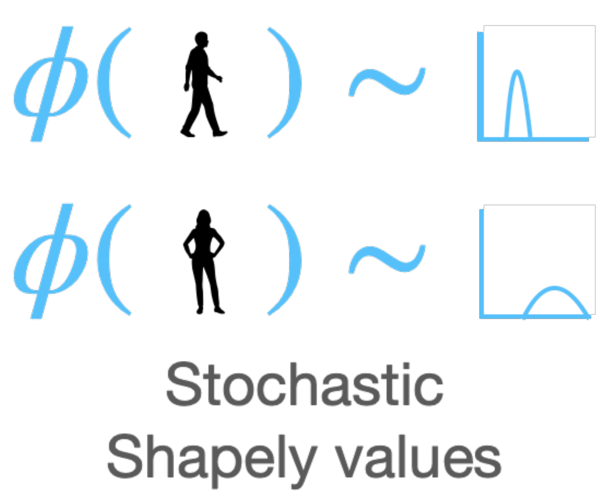
1. We introduce the first GP-specific SHAP algorithm that explain Gaussian processes with Stochastic Shapley values: propagating predictive uncertainty to explanations while preserving their analytical properties.
2. We study the instance-to-explanations regression problem and propose a Shapley GP to predict explanations for new instance, without the need to assess the underlying function to explain.

STOCHASTIC SHAPLEY VALUES



How to split?

Efficient ✅ ?
Symmetric ✅ ?
Null player ✅?



Same formula, but quantities are now random variables

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

can now compute VARIANCE

HOW TO OBTAIN STOCHASTIC EXPLANATIONS FOR GPs?

TL;DR: Conditional expectations of GPs are still GP!

Step 1: Build stochastic game out of posterior GP $f \sim GP(m, \kappa)$

$$\nu_{f,x}(S) := \mathbb{E}_X\left[f(X) \mid X_S = x_S\right] \sim \mathcal{N}(\tilde{m}_S(x_S), \tilde{\kappa}_S(x_S, x_S))$$

where $\tilde{m}_S(x_S) := \mathbb{E}[m(X) \mid X_S = x_S]$ and $\tilde{\kappa}_S(x_S, x_S) := \mathbb{E}[\kappa(X, X') \mid X_S = x_S, X'_S = x_S]$

SVs can be computed through linear operations

Step 2: Rewrite stochastic Shapely value as a weighted linear regression solution

$$\overrightarrow{\phi} = A\mathbf{v}_{f,x}$$

where A is the regression matrix and $\mathbf{v}_{f,x} = [\nu_{f,x}(S_1), \ldots, \nu_{f,x}(S_{2^d})]^\top$ is the vector of game evaluations

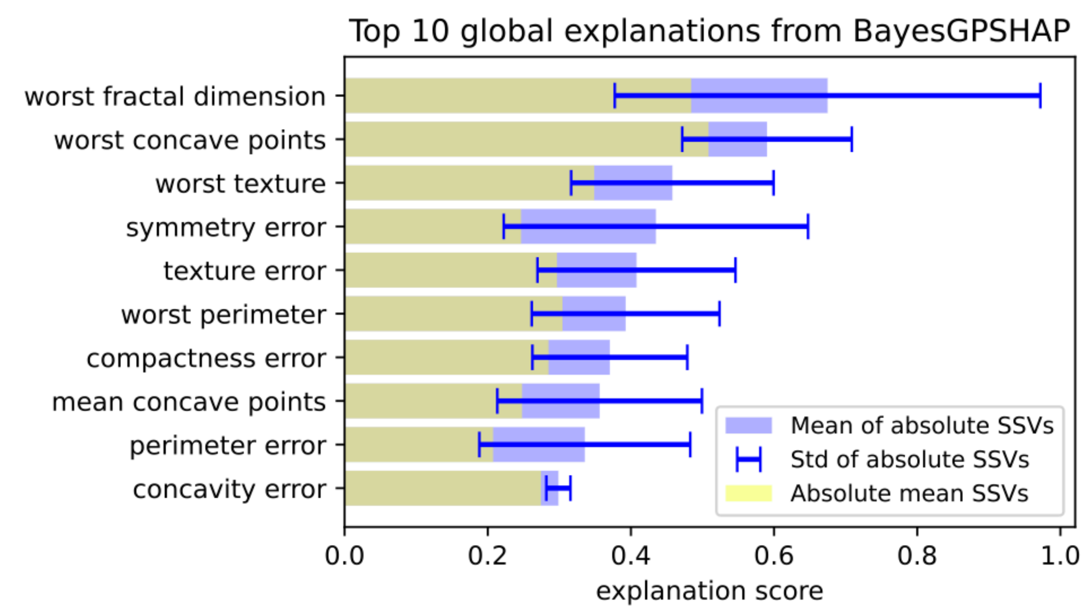
Stochastic SVs for GPs are still Gaussian!

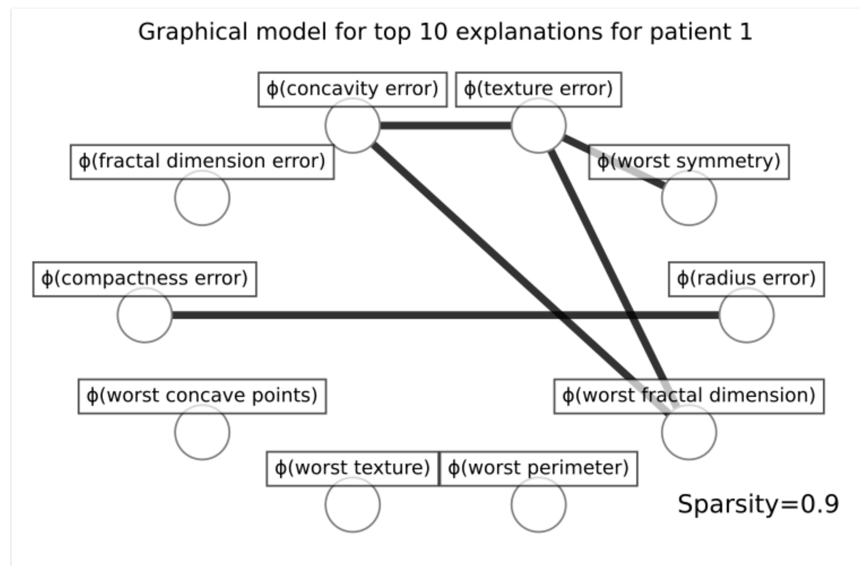
Step 3: Linear operations preserve Gaussianity of the stochastic games from GP

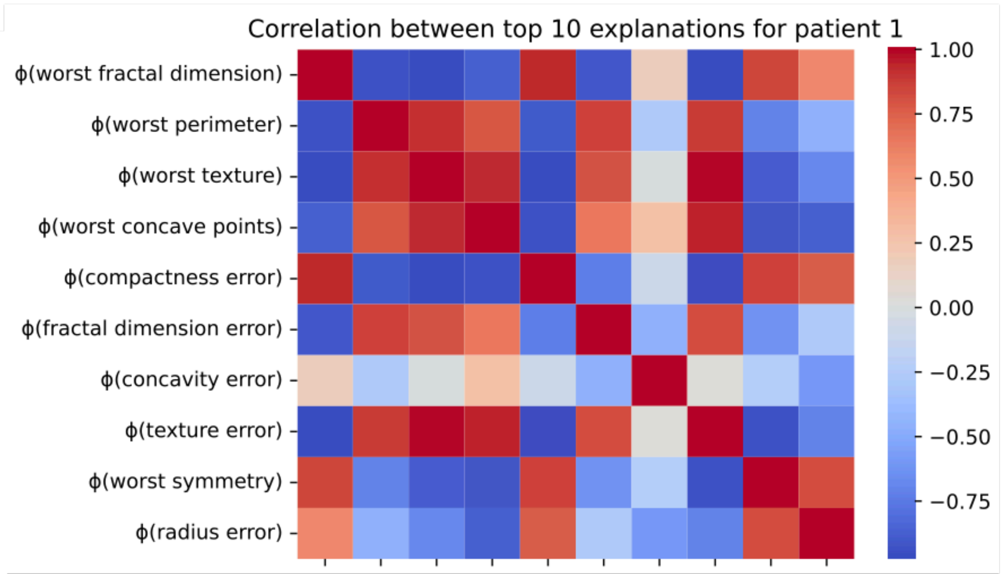
$$\overrightarrow{\phi} \sim \mathcal{N}(A\mathbb{E}[\mathbf{v}_{f,x}], A\mathbb{C}[\mathbf{v}_{f,x}]A^\top)$$

where $\mathbb{E}$ and $\mathbb{C}$ are the expectation and covariance operator

We can now reason about uncertainty, correlation, and independences across explanations!

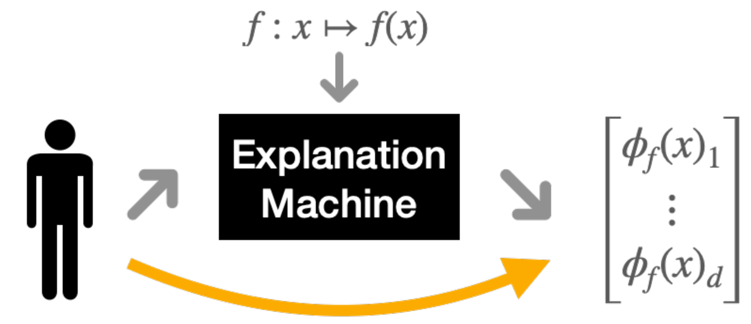






HOW TO PREDICT EXPLANATIONS USING GPs?

Can we bypass the explanation machine?



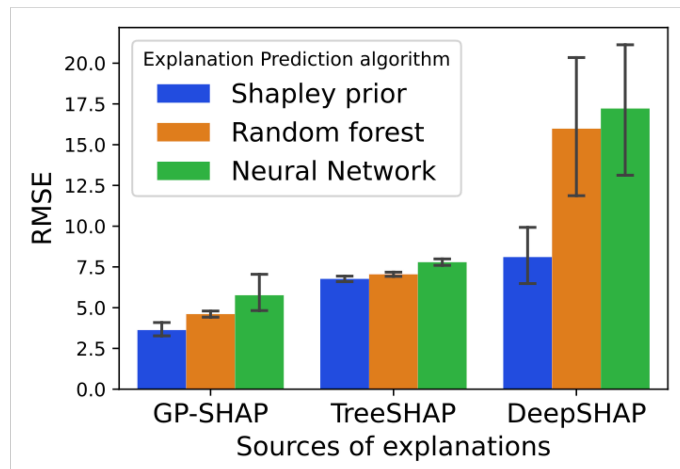
Perform instance-to-explanation regression!

Challenge: Any standard regression would not yield predictions that are Shapley values!

Here comes a Shapley GP with the Shapley prior kernel

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

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



1) train f using GP, Tree, DL
2) Obtain explanations
3) Form regression dataset
4) Predict using either Shapley GP, Random Forest, or Neural network

Paper and code available!