SUPPLEMENTARY MATERIAL: Explaining the uncertain: Stochastic Shapley values for Gaussian process models

A The GP-SHAP algorithm and discussion on computation techniques

We present the complete algorithm for both GP-SHAP and BayesGP-SHAP in Algorithm 1.

Algorithm 1 GP-SHAP / BayesGP-SHAP

Input: Posterior mean function $\tilde{m}$, posterior covariance function $\tilde{k}$, inducing locations $\tilde{\mathbf{X}}$, explanation instances $\mathbf{X}$, number of coalition samples n_Z, hyperparameter λ, n_0, σ_0^2, base kernel k, algorithm **algo**,

1: Compute n_I = number of inducing location, n = number of explanation instances, d = number of features.
2: Compute Cholesky decomposition on posterior covariance $\mathbf{L}\mathbf{L}^\top = \tilde{\mathbf{K}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}$
3: Sample coalitions $\mathcal{S} = \{S_1, ..., S_{n_Z}\}$ from $[d]$, build binary matrix $\mathbf{Z} = \{0,1\}^{n_Z \times d}$ from $\mathcal{S}$, and compute weights $W = \mathrm{diag}[w_1, ..., w_{n_Z}]$ with $w_i = \frac{d-1}{\binom{d}{|S_i|}|S_i|(d-|S_i|)}$.
4: Compute $\mathbf{A} = (\mathbf{Z}^\top W \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{W}$ ▷ Shape: $d \times n_Z$
5: Compute $\mathbf{B}(\mathbf{X}, \mathcal{S}) = [(\mathbf{K}_{\tilde{\mathbf{X}}_S\tilde{\mathbf{X}}_S} + \lambda I)^{-1} k_S(\tilde{\mathbf{X}}_S, \mathbf{X}_S)$ **for** S in $\mathcal{S}]$ ▷ Shape: $n_Z \times n_I \times n$
6: Compute $\mathbf{Q}$ where $\mathbf{Q}_{i,l,k} = \sum_j \mathbf{B}(\mathbf{X}, \mathcal{S})_{i,j,k}\mathbf{L}_{j,l}$ ▷ Shape: $n_Z \times n \times n_I$
7: Compute $\mathbf{R}$ where $\mathbf{R}_{i,k,l} = \sum_j \mathbf{A}_{i,j}\mathbf{Q}_{j,k,l}$ ▷ Shape: $d \times n \times n_I$
8: Compute $\mathbf{V}$ where $\mathbf{V}_{i,m,k,n} = \sum_{j,l} \mathbf{R}_{i,j,k}\mathbf{R}_{m,l,n}$ ▷ Shape: $d \times d \times n \times n$
9: Compute $\mathbf{E}$ where $\mathbf{E}_{i,k} = \sum_j \mathbf{B}(\mathbf{X}, \mathcal{S})_{i,j,k}\tilde{m}(\tilde{\mathbf{X}})_j$ ▷ Shape: $n_Z \times n$
10: Compute $\Phi = \mathbf{A}\mathbf{E}$ ▷ The mean stochastic Shapley values of shape $d \times n$
11: **if algo** = GP-SHAP **then**
12:     **return** mean explanations Φ and covariance $\mathbf{V}$ between d features and n instances
13: **else if algo** = BayesGP-SHAP **then**
14:     Compute $s^2 = \mathrm{diag}\left((\mathbf{E} - \mathbf{Z}\Phi)^\top \mathbf{W}(\mathbf{E} - \mathbf{Z}\Phi)\right) + \mathrm{diag}(\Phi^\top\Phi)$ ▷ Shape: $n \times 1$
15:     Sample σ^2 from Scaled-Inv- $\chi^2\left(n_0 + n_Z, \frac{n_0\sigma_0^2 + n_Z s^2}{n_0 + n_Z}\right)$ ▷ Shape: $n \times 1$
16:     **return** mean explanations Φ and covariance $\mathbf{V} + (\mathbf{Z}^\top \mathbf{W}\mathbf{Z})^{-1}\sigma^2$
17: **end if**

Computational considerations. In terms of computational complexity, one of the most demanding operations in the algorithm is the computation of conditional mean embeddings in step 5. Instead of naively inverting an $n \times n$ matrix, which would have a computational cost of $\mathcal{O}(n^3)$, we employ the conjugate gradient method to reduce the computation of the conditional mean embedding component to $\mathcal{O}(n^2 a)$, where $a \ll n$ represents the number of conjugate gradient iterations. Additionally, to further reduce runtime, we utilize the variational sparse GP model [48]. This model learns a set of inducing locations $\tilde{\mathbf{X}}$ with a size of $n_I \ll n$, which can be reused for the estimation of conditional mean embeddings in the algorithm. Consequently, the computation of the conditional expectation is reduced from $\mathcal{O}(n^2 a)$ to $\mathcal{O}(n_I^2 a)$. Another computational burden arises from the computation of the full covariance matrix across d features and n instances, which requires storage of a $n^2 d^2$ matrix. However, since the full covariance matrix can be factorized into the $\mathbf{R}$ component from step 7 of the algorithm, we can store this low-rank component and compute covariances between specific instances when necessary. It is worth noting that this decomposition of the covariance matrix allows us to avoid redundant computations when computing the covariance component, as we no longer need to iterate over all possible coalitions twice. Finally, we can further speed up our computational by parallelising computation across the sub-sampled coalitions in step 5.

B Proofs and derivations

B.1 Section 2 proofs: Stochastic Shapley values

We include the full proof of the derivation of stochastic Shapley values for completeness. The proof is analogous to the original work of Shapley's [1] but extended to random variable payoffs. Ma et al. [16] has also proved the same theorem but used a different proving strategy. They started with the solution and showed it satisfies the axioms and then prove uniqueness, whereas the following proof starts from the characterisation of s-games and derive the solution from a bottom-up fashion.

To facilitate the proof, we first introduce the concept of stochastic symmetric game.

Proposition 15 (s-symmetric games)**.** *Let C be a real-valued random variables, then the symmetric game $\nu_{C,R}(S) := C\mathbf{1}[R \subseteq S]$ gets a stochastic shapley value as,*

$$\phi_i(\nu_{C,R}) = \frac{C}{r} \tag{16}$$

where $r = |R|$.

Proof. Take any $i, j \in R$, pick a permutation $\pi \in \Pi(U)$ so that $\pi R = R$ and $\pi i = j$, so the induced game $\pi\nu_{C,R} = \nu_{C,R}$, and therefore by the s-symmetry axiom,

$$\phi_j(\nu_{C,R}) = \phi_i(\nu_{C,R}) \tag{17}$$

Now by the s-efficiency axiom,

$$C = \nu_{C,R}(R) = \sum_{j\in R} \phi_j(\nu_{C,R}) = r\phi_i(\nu_{C,R}) \tag{18}$$

for any $i \in R$. □

Now we can characterise the form of any stochastic game as follows:

Proposition 16. *All s-games with finite carrier can be written as a linear combination of s-symmetric games,*

$$\nu = \sum_{R\subseteq N, R\neq\emptyset} \nu_{c_R(\nu),R} \tag{19}$$

where

$$C_R(\nu) = \sum_{T\subseteq R} (-1)^{r-t}\nu(T) \tag{20}$$

Proof. We start by verifying

$$\nu(S) = \sum_{R\subseteq N, R\neq\emptyset} \nu_{c_R(\nu),R}(S) \tag{21}$$

holds for all $S \subseteq U$, and for any finite carrier N of ν. If $S \subseteq N$, then we can rewrite the expression as,

$$\begin{aligned}
\nu(S) &= \sum_{R\subseteq S}\sum_{T\subseteq R} (-1)^{r-t}\nu(T) && (22)\\
&= \sum_{T\subseteq S}\sum_{T\subseteq R\subseteq S} (-1)^{r-t}\nu(T) && (23)\\
&= \sum_{T\subseteq S}\nu(T)\sum_{r=t}^{s}(-1)^{r-t}\binom{s-t}{r-t} && (24)\\
&= \nu(S) && (25)
\end{aligned}$$

where in the last equation we used the fact that $\sum_{r=t}^{s}(-1)^{r-t}\binom{s-t}{r-t}$ is a binomal expansion of $(1+(-1))^{s-t}$, therefore the only non-zero expression is when $t = s$.

□

We can now prove the uniqueness of stochastic Shapley values,

Theorem 4 (Stochastic Shapley values)**.** *The only stochastic value allocation ϕ of ν satisfying s-symmetry, s-efficiency, and s-linearity takes the following form,*

$$\phi_i(\nu) = \sum_{S \subseteq N \setminus \{i\}} c_{|S|} \left(\nu(S \cup i) - \nu(S)\right) \tag{1}$$

where N is the smallest carrier set of Ω, $c_{|S|} = \frac{1}{|N|}\binom{|N|-1}{|S|}^{-1}$ and $\phi_i(\nu)$ is the i^{th} SSV of s-game ν.

Proof. First, let us denote

$$\gamma_i(S) := \sum_{\substack{R \subseteq N \\ S \cup \{i\} \subseteq R}} (-1)^{r-s} \frac{1}{r}.$$

Applying the s-linearity axiom on ϕ to the characterisation of ν from the previous propositions leads us to the following,

$$\begin{aligned}
\phi_i(\nu) &= \phi_i \left(\sum_{R \subseteq N, R \neq \emptyset} \nu_{C_R(\nu), R} \right) && (26) \\
&= \sum_{R \subseteq N, R \neq \emptyset} \phi_i(\nu_{C_R(\nu), R}) && (27) \\
&= \sum_{R \subseteq N, i \in R} c_R(\nu) \frac{1}{r} && (28) \\
&= \sum_{R \subseteq N, i \in R} \frac{1}{r} \left(\sum_{S \subseteq R} (-1)^{r-s} \nu(S) \right) && (29) \\
&= \sum_{S \subseteq N} \sum_{\substack{R \subseteq N \\ S \cup \{i\} \subseteq R}} (-1)^{r-s} \nu(S) \frac{1}{r} && (30) \\
&= \sum_{S \subseteq N} \gamma_i(S) \nu(S) && (31) \\
&= \sum_{\substack{S \subseteq N \\ i \in S}} \gamma_i(S)\nu(S) + \gamma_i(S - \{i\})\nu(S - \{i\}) && (32) \\
&= \sum_{\substack{S \subseteq N \\ i \in S}} \gamma_i(S) \left(\nu(S) - \nu(S - \{i\})\right) && (33) \\
&= \sum_{\substack{S \subseteq N \\ i \in S}} \frac{(s-1)!(n-s)!}{n!} \left(\nu(S) - \nu(S - \{i\})\right) && (34) \\
&= \sum_{S \subseteq N \setminus \{i\}} c_{|S|} \left(\nu(S \cup i) - \nu(S)\right) && (35)
\end{aligned}$$

where in (32) we used the following observation: given $i \notin S' \subseteq N$, and $S = S' \cup \{i\}$, then $\gamma_i(S) = -\gamma_i(S')$.

It satisfies uniqueness by construction. □

Proposition 5. *Given the player set Ω, let ν be a stochastic game, ϕ a stochastic Shapley value allocation, and $\bar{\phi}$ a deterministic Shapley value allocation. Suppose that $\mathbb{E}[\nu]$ and $\mathbb{V}[\nu]$ are the corresponding mean and variance d-games, respectively. Then, $\mathbb{E}[\phi(\nu)] = \bar{\phi}(\mathbb{E}[\nu])$, but $\mathbb{V}[\phi(\nu)] \neq \bar{\phi}(\mathbb{V}[\nu])$. In particular, the SSV variance is given by*

$$\mathbb{V}[\phi_i(\nu)] = \sum_{S \subseteq N \setminus \{i\}} \sum_{S' \subseteq N \setminus \{i\}} c_{|S|} c_{|S'|} \big(\mathbb{C}[\nu_{S \cup i}, \nu_{S' \cup i}] - \mathbb{C}[\nu_{S \cup i}, \nu_{S'}] - \mathbb{C}[\nu_S, \nu_{S' \cup i}] + \mathbb{C}[\nu_S, \nu_{S'}] \big),$$

where $\nu_S = \nu(S)$ and $\mathbb{C}$ is the covariance function between the stochastic payoffs.

Proof. The equivalence between mean of stochastic Shapley values and deterministic Shapley values of mean game is trivial to show leveraging the linearity of expectation. The variance of $\mathbb{V}[\phi_i(\nu)]$ can be shown by repeatedly applying the standard identity $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\mathbb{C}[X,Y]$ for random variables X, Y. Now consider the deterministic Shapley values of variance game $\mathbb{V}[\nu]$,

$$\bar{\phi}_i[\mathbb{V}[\nu(\cdot)]] = \sum_{S\subseteq N\setminus\{i\}} c_{|S|} \left(\mathbb{V}[\nu(S\cup i)] - \mathbb{V}[\nu(S)]\right) \tag{36}$$

Comparing to the expression of $\mathbb{V}[\phi_i(\nu)]$ from the lemma,

$$\mathbb{V}[\phi_i(\nu)] = \sum_{S\subseteq N\setminus\{i\}} \sum_{S'\subseteq N\setminus\{i\}} c_{|S|}c_{|S'|}\big(\mathbb{C}[\nu_{S\cup i}, \nu_{S'\cup i}] - \mathbb{C}[\nu_{S\cup i}, \nu_{S'}] - \mathbb{C}[\nu_S, \nu_{S'\cup i}] + \mathbb{C}[\nu_S, \nu_{S'}]\big),$$

even if we assume mutual independence across all payoff random variables, leading to $\mathbb{C}[\nu(S \cup i), \nu(S)] = 0$ for all S, we still would not subtract but instead sum the variance of $\mathbb{V}[\nu(S\cup i)]$ and $\mathbb{V}[\nu(S)]$. Therefore the variances of stochastic Shapley values is not the same as the deterministic Shapley values of the variance game. □

B.2 Section 3.1 proofs on the stochastic Shapley values for induced stochastic game from GP

Proposition 6 (Stochastic game ν_f as induced GP)**.** *Let $f \sim \mathcal{GP}(\tilde{m}, \tilde{k})$ with integrable sample paths, i.e. $\int_{\mathcal{X}} |f| dp_X < \infty$ almost surely. The stochastic payoff function ν_f induced by f is a Gaussian process with the following mean and covariance functions:*

$$m_\nu(\mathbf{x}, S) := \mathbb{E}_X[\tilde{m}(X) \mid X_S = \mathbf{x}_S], \tag{4}$$

$$k_\nu\left((\mathbf{x}, S), (\mathbf{x}', S')\right) := \mathbb{E}_{X,X'}\left[\tilde{k}(X, X') \mid X_S = \mathbf{x}_S, X'_{S'} = \mathbf{x}'_{S'}\right]. \tag{5}$$

Proof. This is a direct application of Chau et al. [18, Proposition 3.2] to the distribution $P(X \mid X_S = \mathbf{x}_S)$. □

Theorem 7 (Stochastic Shapley values of ν_f)**.** *Let ν_f be an induced stochastic game from the GP $f \sim \mathcal{GP}(\tilde{m}, \tilde{k})$ and denote $\mathbf{v}_\mathbf{x} := [\nu_f(\mathbf{x}, S_1), \ldots \nu_f(\mathbf{x}, S_{2^d})]^\top$ 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,*

$$\phi(\nu_f(\mathbf{x}, \cdot)) \sim \mathcal{N}(\mathbf{A}\mathbb{E}[\mathbf{v}_\mathbf{x}], \mathbf{A}\mathbb{V}[\mathbf{v}_x]\mathbf{A}^\top) \quad \textit{with} \quad \mathbf{A} := (\mathbf{Z}^\top\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{W}, \tag{6}$$

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.

Proof. Recall from Lundberg and Lee [2, Theorem 2], for deterministic Shapley values, given a deterministic payoff $\bar{\mathbf{v}}_\mathbf{x}$ for all 2^d coalitions, the expression of Shapley values for each $i \in [d]$,

$$\bar{\phi}_{\mathbf{x}i} = \sum_{S\subseteq[d]\setminus\{i\}} c_{|S|} \left(\bar{\nu}_f(S\cup i) - \bar{\nu}_f(S)\right) \tag{37}$$

can be written compactly as the following vector,

$$\bar{\phi}_\mathbf{x} = \mathbf{A}\bar{\mathbf{v}}_\mathbf{x}. \tag{38}$$

We can therefore similarly write down the form of the stochastic Shapley values using this linear operator $\mathbf{A}$, acting now on a vector of random variable output stochastic payoff vector $\mathbf{v}_\mathbf{x}$,

$$\phi_\mathbf{x} = \mathbf{A}\mathbf{v}_\mathbf{x}. \tag{39}$$

Nonetheless, as Proposition 8 implies that $\mathbf{v}_\mathbf{x}$ is a multivariate Gaussian, therefore $\phi_\mathbf{x}$ is also multivariate Gaussian with mean and covariance the following,

$$\mathbf{v}_\mathbf{x} \sim \mathcal{N}\left(A\mathbb{E}[\mathbf{v}_\mathbf{x}], A\mathbb{V}[\mathbf{v}_\mathbf{x}]A^\top\right). \tag{40}$$

□

B.3 Section 3.2 proofs on estimation

To proceed, we first introduce the concepts of conditional mean embedding as a tool to estimate conditional expectation of functions living in their corresponding RKHSs,

Definition 17 (Conditional mean embedding [38])**.** *Let X, Y be random variables and $k : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$ a kernel on X, then we define the following as the conditional mean embedding of $p(X \mid Y = y)$,*

$$\mu_{X|Y=y} := \int k(\cdot, X) d\mathbb{P}(X \mid Y = y) \tag{41}$$

Proposition 18 (Conditional Mean estimation)**.** *For random variable X, Y, and a kernel $k : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$ on $\mathcal{X}$ and a kernel $l : \mathcal{Y} \to \mathcal{Y} \to \mathbb{R}$ on $\mathcal{Y}$. Given observations $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\}$, the empirical conditional mean embedding can be estimated as*

$$\hat{\mu}_{X|Y=y} = l(y, \mathbf{y}) \left(\mathbf{L}_{\mathbf{yy}} + \lambda I\right)^{-1} k(\mathbf{X}, \cdot), \tag{42}$$

where $l(y, \mathbf{y}) = [l(y, y_1), \ldots, l(y, y_n)]^\top$ and $k(\cdot, \mathbf{X}) = [k(\cdot, \mathbf{x}_1), \ldots, k(\cdot, \mathbf{x}_n)]^\top$, the parameter $\lambda > 0$ is there to stablise the inversion. Now for $f \in \mathcal{H}_k$, the conditional expectation can then be estimated as,

$$\begin{aligned} \hat{\mathbb{E}}[f(X) \mid Y = y] &= \langle \hat{\mu}_{X|Y=y}, f \rangle && (43) \\ &= l(y, \mathbf{y})(\mathbf{L}_{\mathbf{yy}} + \lambda I)^{-1} \mathbf{f}, && (44) \end{aligned}$$

where $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top$.

Proof. This is standard result from literature, please read Song et al. [49], Muandet et al. [38] for more details. □

Now we can apply these propositions to estimate the mean and covariance functions of the induced stochastic game from GP,

Proposition 8 (Estimating ν_f)**.** *Given $\mathbf{D} = (\mathbf{X}, \mathbf{y})$ and the posterior GP $f \mid \mathbf{D} \sim \mathcal{GP}(\tilde{m}, \tilde{k})$, the mean and covariance function of the stochastic cooperative game ν_f can be estimated as,*

$$\hat{m}_\nu(\mathbf{x}, S) = \mathbf{b}(\mathbf{x}, S)^\top \tilde{m}(\mathbf{X}), \qquad \hat{k}_\nu\left((\mathbf{x}, S), (\mathbf{x}', S')\right) = \mathbf{b}(\mathbf{x}, S)^\top \tilde{\mathbf{K}}_{\mathbf{XX}} \mathbf{b}(\mathbf{x}', S'), \tag{7}$$

where $\mathbf{b}(\mathbf{x}, S) := (\mathbf{K}_{\mathbf{X}_S\mathbf{X}_S} + \lambda I)^{-1} k_S(\mathbf{X}_S, \mathbf{x}_S)$, $\tilde{m}(\mathbf{X}) = [\tilde{m}(\mathbf{x}_1), \ldots, \tilde{m}(\mathbf{x}_n)]^\top$, and $k_S : \mathcal{X}_S \times \mathcal{X}_S \to \mathbb{R}$ is the kernel defined on the sub-feature space of $\mathcal{X}$ and we write $k_S(\mathbf{x}_S, \mathbf{X}_S) := [k_S(\mathbf{x}_S, \mathbf{x}_{1S}), ..., k_S(\mathbf{x}_S, \mathbf{x}_{nS})]$ and $\mathbf{K}_{\mathbf{XX}}$ and $\tilde{\mathbf{K}}_{\mathbf{XX}}$ as the gram matrix of $\mathbf{X}$ using kernel k and $\tilde{k}$ respectively. The parameter $\lambda > 0$ is a fixed hyperparameter to stabilise the inversion.

Proof. Without loss of generality, we will demonstrate this proposition with $\tilde{m}, \tilde{k}$ obtained via standard GP regression, i.e.,

$$\begin{aligned} \tilde{m}(\mathbf{x}) &= k(\mathbf{x}, \mathbf{X})(\mathbf{K}_{\mathbf{XX}} + \sigma^2 I)^{-1} \mathbf{y} && (45) \\ \tilde{k}(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X})(\mathbf{K}_{\mathbf{XX}} + \sigma^2)^{-1} k(\mathbf{X}, \mathbf{x}'). && (46) \end{aligned}$$

Starting with the mean function,

$$\begin{aligned} \mathbb{E}[\tilde{m}(X) \mid X_S = \mathbf{x}_S] &= \mathbb{E}_X[k(X, \mathbf{X})(\mathbf{K}_{\mathbf{XX}} + \sigma^2 I)^{-1} \mathbf{y} \mid X_S = \mathbf{x}_S] && (47) \\ &= \langle k(\cdot, \mathbf{X})(\mathbf{K}_{\mathbf{XX}} + \sigma^2 I)^{-1} \mathbf{y}, \mu_{X|X_S=\mathbf{x}_S} \rangle_{\mathcal{H}_k}. && (48) \end{aligned}$$

We can replace the population conditional mean embedding with the empirical version, and expand,

$$\begin{aligned} \hat{\mathbb{E}}[\tilde{m}(X) \mid X_S = \mathbf{x}_S] &= \langle k(\cdot, \mathbf{X})(\mathbf{K}_{\mathbf{XX}} + \sigma^2 I)^{-1} \mathbf{y}, \hat{\mu}_{X|X_S=\mathbf{x}_S} \rangle_{\mathcal{H}_k} && (49) \\ &= k_S(\mathbf{X}_S, \mathbf{x}_S)(\mathbf{K}_{\mathbf{X}_S\mathbf{X}_S} + \lambda I)^{-1} \mathbf{K}_{\mathbf{XX}}(\mathbf{K}_{\mathbf{XX}} + \sigma^2 I)^{-1} \mathbf{y} && (50) \\ &= \mathbf{b}(\mathbf{x}, S)^\top \tilde{m}(\mathbf{X}). && (51) \end{aligned}$$

Analogously, the conditional expectation of the posterior covariance function, i.e., $\mathbb{E}[\tilde{k}(X, X') \mid X_S = \mathbf{x}_S, X'_S = \mathbf{x}'_S]$, can be estimated following the steps above,

$$\mu_{X|X_S=\mathbf{x}_S}^\top \mu_{X'|X'_S=\mathbf{x}'_S} - \mu_{X|X_S=\mathbf{x}_S}^\top k(\cdot, \mathbf{X})(\mathbf{K}_{\mathbf{XX}} + \sigma^2 I)^{-1} k(\mathbf{X}, \cdot) \mu_{X'|X'_S=\mathbf{x}'_S}. \tag{52}$$

After replacing the population conditional mean embedding as their empirical estimates, we can arrive at the solution. □

Proposition 9 (**GP-SHAP**)**.** *Let the matrix* $\mathbf{A}$ *be defined as in Theorem 7. The mean and covariance for the multivariate stochastic Shapley values can be estimated as,*

$$\phi\left(\hat{\nu}_f(\mathbf{x},\cdot)\right) = \mathcal{N}\left(\mathbf{A}\mathbf{B}(\mathbf{x},[d])^\top \tilde{m}(\mathbf{X}), \mathbf{A}\mathbf{B}(\mathbf{x},[d])^\top \tilde{\mathbf{K}}_{\mathbf{X}\mathbf{X}}\mathbf{B}(\mathbf{x},[d])\mathbf{A}^\top\right) \tag{8}$$

where $\mathbf{B}(\mathbf{x},[d]) = [\mathbf{b}(\mathbf{x},[d]_1),\dots,\mathbf{b}(\mathbf{x},[d]_{2^d})]^\top$.

Proof. The result follows directly from the previous proposition. Recall $\phi(\hat{\nu}_f(\mathbf{x},\cdot)) = \mathbf{A}\hat{\mathbf{v}}_{\mathbf{x}}$ for $\hat{\mathbf{v}}_{\mathbf{x}}$ the vector of stochastic payoffs for each coalition. To estimate the mean, we

$$\begin{aligned}
\mathbb{E}[\phi(\hat{\nu}_f(\mathbf{x},\cdot))] &= \mathbf{A}\mathbb{E}[\hat{\mathbf{v}}_{\mathbf{x}}] && (53)\\
&= \mathbf{A}\begin{bmatrix}\hat{m}_\nu(\mathbf{x},S_1)\\ \vdots \\ \hat{m}_\nu(\mathbf{x},S_{2^d})\end{bmatrix} && (54)\\
&= \mathbf{A}\begin{bmatrix}\mathbf{b}(\mathbf{x},S_1)^\top\tilde{m}(\mathbf{X})\\ \vdots \\ \mathbf{b}(\mathbf{x},S_{2^d})^\top\tilde{m}(\mathbf{X})\end{bmatrix} && (55)\\
&= \mathbf{A}\mathbf{B}(\mathbf{x},[d])^\top\tilde{m}(\mathbf{X}). && (56)
\end{aligned}$$

Recall $\mathbb{V}[\mathbf{v}_{\mathbf{x}}]_{i,j} = \hat{k}_\nu((\mathbf{x},S_i),(\mathbf{x},S_j)) = \mathbf{b}(\mathbf{x},S_i)^\top\tilde{\mathbf{K}}_{\mathbf{X}\mathbf{X}}\mathbf{b}(\mathbf{x},S_j)$, the derivation for the covariance matrix then follows analogously as the derivation for the mean,

$$\begin{aligned}
\mathbb{V}[\phi(\hat{\nu}_f(\mathbf{x},\cdot))] &= \mathbf{A}\mathbb{V}[\hat{\mathbf{v}}_{\mathbf{x}}]\mathbf{A}^\top && (57)\\
&= \mathbf{A}\left[\mathbf{b}(\mathbf{x},S_i)^\top\tilde{\mathbf{K}}_{\mathbf{X}\mathbf{X}}\mathbf{b}(\mathbf{x},S_j)\right]_{i=1,j=1}^{2^d,2^d}\mathbf{A}^\top && (58)\\
&= \mathbf{A}\mathbf{B}(\mathbf{x},[d])^\top\tilde{\mathbf{K}}_{\mathbf{X}\mathbf{X}}\mathbf{B}(\mathbf{x},[d])\mathbf{A}^\top. && (59)
\end{aligned}$$

□

Proposition 10 (BayesSHAP [20])**.** *Given the data generation above, the posterior distribution on* $\bar{\phi}$ *and* σ^2 *follows:*

$$\bar{\phi} \mid \sigma^2, \mathbf{Z}_\ell, f, \mathbf{x}, \mathbf{D} \sim \mathcal{N}(\mathbf{A}_\ell\bar{\mathbf{v}}_{\mathbf{x}}, (\mathbf{Z}_\ell^\top\mathbf{W}_\ell\mathbf{Z}_\ell)^{-1}\sigma^2) \tag{11}$$

$$\sigma^2 \mid \mathbf{Z}_\ell, f, \mathbf{x}, \mathbf{D} \sim \text{Scaled-Inv-}\chi^2\left(\ell_0+\ell, \frac{\ell_0\sigma_0^2 + \ell s^2(\bar{\mathbf{v}}_{\mathbf{x}})}{\ell_0+\ell}\right) \tag{12}$$

where ℓ *is the number of coalitions* $\boldsymbol{\mathcal{S}} = \{S_j\}_{j=1}^\ell$ *we sample uniformly from* $2^{[d]}$, $\mathbf{Z}_\ell$ *is the binary matrix representing* $\boldsymbol{\mathcal{S}}$, *and* $\mathbf{W}_\ell$ *is the corresponding weight matrix, and* $\mathbf{A}_\ell = (\mathbf{Z}_\ell^\top\mathbf{W}_\ell\mathbf{Z}_\ell)^{-1}\mathbf{Z}_\ell^\top\mathbf{W}_\ell$ *is the WLS matrix,* $\bar{\mathbf{v}}_{\mathbf{x}} = [\bar{\nu}_f(\mathbf{x},S_1),...,\bar{\nu}_f(\mathbf{x},S_\ell)]^\top$ *is the vector of deterministic payoffs, and*

$$s^2(\bar{\mathbf{v}}_{\mathbf{x}}) = \frac{1}{\ell}\left[(\bar{\mathbf{v}}_{\mathbf{x}} - \mathbf{Z}_\ell\mathbf{A}_\ell\bar{\mathbf{v}}_{\mathbf{x}})^\top W_\ell(\bar{\mathbf{v}}_{\mathbf{x}} - \mathbf{Z}_\ell\mathbf{A}_\ell\bar{\mathbf{v}}_{\mathbf{x}}) + (\mathbf{A}_\ell\bar{\mathbf{v}}_{\mathbf{x}})^\top(\mathbf{A}_\ell\bar{\mathbf{v}}_{\mathbf{x}})\right] \tag{13}$$

measures the average weighted error in the regression and the norm of the mean explanations.

Proof. See Slack et al. [20, Section. 3.1]. □

Proposition 11 (**BayesGP-SHAP**)**.** *Continuing from Propositions 9 and 10, the posterior distribution of the stochastic Shapley values can be estimated using the Bayesian WLS approach as,*

$$\phi \mid \sigma^2, \mathbf{Z}_\ell, \mathbf{x}, \mathbf{D} \sim \mathcal{N}\left(\mathbf{A}_\ell\mathbf{B}(\mathbf{x},\boldsymbol{\mathcal{S}}))^\top\tilde{m}(\mathbf{X}), \mathbf{A}_\ell\mathbf{B}(\mathbf{x},\boldsymbol{\mathcal{S}})^\top\tilde{\mathbf{K}}_{\mathbf{X}\mathbf{X}}\mathbf{B}(\mathbf{x},\boldsymbol{\mathcal{S}})\mathbf{A}_\ell^\top + (\mathbf{Z}_\ell^\top\mathbf{W}_\ell\mathbf{Z}_\ell)^{-1}\sigma^2\right)$$

where σ^2 *is sampled from* $\sigma^2 \mid \mathbf{Z}_\ell \sim \text{Scaled-Inv-}\chi^2\left(\ell_0+\ell, \frac{\ell_0\sigma_0^2+\ell s^2(\mathbb{E}[\mathbf{v}_{\mathbf{x}}])}{\ell_0+\ell}\right)$.

Proof. We drop the bar notation of $\bar{\phi}$ to unify notations. Given the posterior GP $f \mid \mathbf{D} \sim \mathcal{GP}(\tilde{m},\tilde{k})$

$$p(\phi \mid \sigma^2, \mathbf{Z}_\ell, \mathbf{x}, \mathbf{D}) = \int p(\phi \mid \sigma^2, \mathbf{Z}_\ell, f, \mathbf{x}, \mathbf{D})p(f \mid \mathbf{D})df \tag{60}$$

Using a standard Gaussian conjugacy procedure, we can derive the variance as the sum of variances from GP-SHAP and BayesSHAP. While it is possible to integrate $p(\sigma^2 \mid \mathbf{Z}_\ell, f, \mathbf{x}, \mathbf{D})$ with respect to the posterior, this leads to a complex scaled mixture of normals that is difficult to model. Instead, we construct a scaled inverse chi-square distribution with $s^2[\mathbb{E}[\mathbf{b_x}]]$, which represents the error of the weighted regression with respect to the mean payoffs $\mathbb{E}[\mathbf{v_x}]$. We sample σ^2 from the following distribution:

$$\sigma^2 \mid \mathbf{Z}_\ell, \mathbf{x}, \mathbf{D} \sim \text{Scale-Inv-}\chi^2\left(\ell_0 + \ell, \frac{\ell_0\sigma_0^2 + \ell s^2(\mathbb{E}[\mathbf{v_x}])}{\ell_0 + \ell}\right). \tag{61}$$

$\square$

B.4 Proofs for section 4 on predictive explanation and Shapley prior

Proposition 12 (The Shapley prior over ϕ)**.** *The prior $f \sim \mathcal{GP}(0, k)$ and the corresponding stochastic 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*

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathcal{A}(\mathbf{x})^\top \mathcal{A}(\mathbf{x}'), \quad \mathcal{A}(\mathbf{x}) = \Psi(\mathbf{x})\mathbf{A}^\top \tag{14}$$

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

Proof. The proof is similar to how we proved previous propositions but applied to prior GP $f \sim \mathcal{GP}(0, k)$ instead. If we set,

$$\nu_f(\mathbf{x}, S) = \mathbb{E}[f(X) \mid X_S = \mathbf{x}_S], \tag{62}$$

then ν_f is a GP on the joint space of data and coalitions with mean 0, and covariance function,

$$\begin{aligned} \operatorname{cov}\left(\nu_f(\mathbf{x}, S), \nu_f(\mathbf{x}', S')\right) &= \mathbb{E}[k(X, X') \mid X_S = \mathbf{x}_S, X'_{S'} = \mathbf{x}'_{S'}] &(63)\\ &= \mu_{X\mid X_S = \mathbf{x}_S}^\top \mu_{X\mid X_{S'} = \mathbf{x}'_{S'}}. &(64) \end{aligned}$$

Since $\phi = \mathbf{A}\mathbf{v_x}$ for $\mathbf{v_x}$ the vector of stochastic payoff from the game induced by the GP prior, the mean stays 0, and the covariance is,

$$\begin{aligned} \kappa(\mathbf{x}, \mathbf{x}') &= \mathbf{A}\left[\mu_{X\mid X_{S_i} = \mathbf{x}_{S_i}}^\top \mu_{X\mid X_{S_j} = \mathbf{x}'_{S_j}}\right]_{i=1,j=1}^{2^d, 2^d} \mathbf{A}^\top &(65)\\ &= \mathbf{A}\Psi(\mathbf{x})^\top \Psi(\mathbf{x}')\mathbf{A}^\top &(66)\\ &= \mathcal{A}(\mathbf{x})^\top \mathcal{A}(\mathbf{x}'), &(67) \end{aligned}$$

therefore we have a matrix-valued covariance kernel κ to build a prior over the induced Shapley values. $\square$

Proposition 13 (Predictive explanations as multi-output GPs)**.** *Given $\mathbf{D}_\phi = \{(\mathbf{x}_i, \boldsymbol{\phi}_i)\}_{i=1}^n = (\mathbf{X}, \mathbf{\Phi_X})$ where $\boldsymbol{\phi}_i \in \mathbb{R}^d$ are the Shapley values computed under predictive model f and $\mathbf{\Phi_X} = [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_n]^\top$, the predictive explanations for new data $\mathbf{x}'$ is distributed as,*

$$\phi(\mathbf{x}') \mid \mathbf{D}_\phi \sim \mathcal{N}\left(\tilde{m}_\phi(\mathbf{x}'), \quad \kappa(\mathbf{x}', \mathbf{x}') - \kappa(\mathbf{x}', \mathbf{X})b_\kappa(\mathbf{x}', \mathbf{X})\right) \tag{15}$$

where $\tilde{m}_\phi(\mathbf{x}') = b_\kappa(\mathbf{x}', \mathbf{X})^\top \operatorname{vec}(\Phi_\mathbf{X})$, $b_\kappa(\mathbf{x}', \mathbf{X}) := (\mathcal{K}_\mathbf{XX} + \sigma_\phi^2 I)^{-1}\kappa(\mathbf{X}, \mathbf{x}')$, $\mathcal{K}_\mathbf{XX}$ is the gram matrix for kernel κ of size $nd \times nd$, $\kappa(\mathbf{x}', \mathbf{X}) = [\kappa(\mathbf{x}', \mathbf{x}_1), \ldots, \kappa(\mathbf{x}', \mathbf{x}_n)]$ is of size $d \times nd$ and σ_ϕ^2 is the noise parameter for regression.

Proof. Follows from standard vector-valued Gaussian process regression results. See Alvarez et al. [50] for a detailed discussion on regression with matrix-valued kernels. $\square$

Proposition 14 (Posterior mean as Shapley values for payoff vector $\tilde{\mathbf{v}}_{\mathbf{x}'}$)**.** *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 $(\mathcal{K}_\mathbf{XX} + \sigma_\phi^2 I)^{-1} \operatorname{vec}(\mathbf{\Phi_X})$.*

Proof. There are two ways to see this. First is by brute force and rearranging the terms in the posterior mean expression. The other is to leverage the vector-valued representer theorem [51] and write the posterior mean as,

$$\tilde{m}_\phi(\mathbf{x}') = \sum_{i=1}^n \mathcal{A}(\mathbf{x}')^\top \mathcal{A}(\mathbf{x}_i)\alpha_i, \quad \alpha_i \in \mathbb{R}^d \tag{68}$$

$$= \sum_{i=1}^n \mathbf{A}\Psi(\mathbf{x}')^\top \Psi(\mathbf{x}_i)\mathbf{A}^\top \alpha_i \tag{69}$$

$$= \mathbf{A}\left(\sum_{i=1}^n \Psi(\mathbf{x}')^\top \Psi(\mathbf{x}_i)\mathbf{A}^\top \alpha_i\right) \tag{70}$$

$$= \mathbf{A}\tilde{\mathbf{v}}_{\mathbf{x}'} \tag{71}$$

after some linear algebra exercises, we can see that α_i is the $[i : i + (d-1)]$ sub-vector of $(\mathscr{K}_{\mathbf{XX}} + \sigma_\phi^2 I)^{-1}\operatorname{vec}(\mathbf{\Phi}_\mathbf{X})$ □

C Implementation details and further illustrations.

All illustrations are run locally on a MacbookPro 2021 with Apple M1 pro chip.

C.1 Ablation study on different notions of uncertainties captured

To demonstrate the difference between the uncertainties captured by GP-SHAP, BayesSHAP, and BayesGP-SHAP, we utilise the California housing dataset [41]. This dataset was derived from the 1990 U.S. census, each observation represent a census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). The dataset includes 20640 instances with 8 numerical features measuring the following:

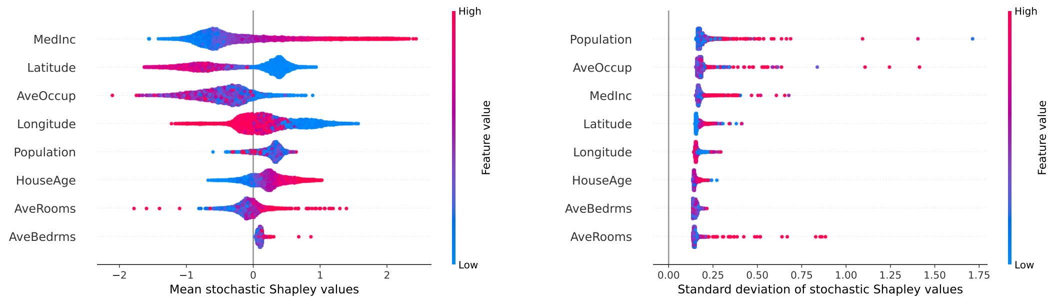
- **MedInc:** Median income in block group
- **HouseAge:** Median house age in block group
- **AveRooms:** Average number of rooms per household
- **AveBedrms:** Average number of bedrooms per household
- **Population:** Block group population
- **AveOccup:** Average number of houehold members
- **Latitude:** Block group latitude
- **Longitude:** Block group longitude

The target variable is the median house value for California districts, expressed in hundreds of thousands of dollars. In the following, we train a GP model and extract explanations using GP-SHAP, BayesSHAP, and BayesGP-SHAP, for 4 different configurations:

1. trained on 25% of data, estimate the Shapley values using 50% of coalitions.
2. trained on 25% of data, estimate the Shapley values using 100% of coalitions.
3. trained on 100% of data, estimate the Shapley values using 50% of coalitions.
4. trained on 100% of data, estimate the Shapley values using 100% of coalitions.

To fit the GP model, we employ a sparse Variational GP approach with 200 learnable inducing point locations. The evidence lower bound is optimized using batch gradient descent with a batch size of 64, a learning rate of 0.01, and 100 iterations. The RBF kernel with learnable bandwidths initialized using the median heuristic approach is used for the sparse GP. The inducing locations are initialized using a standard clustering approach to obtain a representative set of inducing points.

After training the model, we reuse the learned inducing points and kernel bandwidths for the explanation algorithms. The explanations are obtained using the procedure described in Algorithm 1 of our work.



(a) Mean of each stochastic explanations

(b) Standard deviation of each explanations

Figure 4: We plot the beeswarm plot of the mean and standard deviations of each stochastic explanations from BayesGP-SHAP fitted on the housing dataset. The features are ranked according to the distance span by the largest and smallest mean (std) stochastic Shapley values.

In Figure 1 of our paper, we present the stochastic Shapley values for the 11th observation, computed using the three explanation algorithms. The plot includes the 95% credible interval to visualize the uncertainties associated with the explanations.

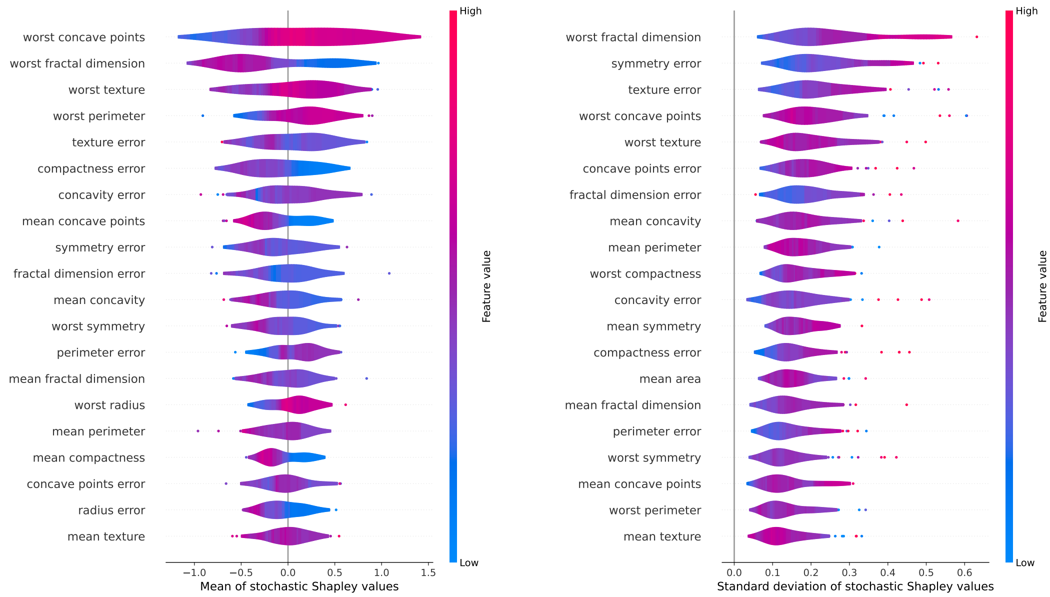
Further illustration: In Figure 4, we plot the beeswarm plot on the mean and standard deviation of each stochastic explanations respectively. We color the point based on the relative size of the feature value compared to the rest. We see that in Figure 4a, which plotted the mean stochastic shapley values for each observation, the relationship between most features' explanation to the target variable is quite linear. For example, the higher the median income (**MedInc**), the more positive those feature contribute to predicting the respective median house value. On the other hand, Figure 4b illustrated the standard deviation of each stochastic explanations. In general, we see that the larger the feature values are, the more uncertain the explanation becomes. Nonetheless, we see that the feature contributing the most, defined as the feature having largest distance spanned by their most positive and most negative mean stochastic Shapley values, does not necessarily have the largest variation respectively.

C.2 Exploratory analysis of the stochastic explanations

For this illustration, we utilise the breast cancer dataset [42], containing 569 patients with 30 numeric features. They are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass and describe characteristics of the cell nuclei present in the image:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness $(\frac{\text{perimeter}^2}{\text{area}-1})$
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The goal is to predict whether a tumour is malignant or benign. We first fit a GP model with RBF kernel using again the sparse Variational GP formulation with 200 learnable inducing locations. We initialise the inducing points using standard clustering techniques on the data. The evidence lower bound objective is optimised with a learning rate of $1e^{-4}$ and 1000 iterations using batch gradient descent of batch size 64. To obtain the explanations, we run the BayesGP-SHAP algorithm with 2^{16}



(a) Mean of each stochastic explanations

(b) Standard deviation of each explanations

Figure 5: We plot the violin plot of the mean and standard deviations of each stochastic explanations from BayesGP-SHAP fitted on the breast cancer. The features are ranked according to the distance span by the largest and smallest mean (std) stochastic Shapley values.

number of coalitions. We do not compare GP-SHAP and BayesSHAP here because the BayesSHAP uncertainties have shrunk to almost zero, i.e., the mean standard deviations from the BayesSHAP uncertainties across all features and data is 0.0002. This reconfirms the fact from Slack et al. [20] that as we increase the sample size the estimation error goes to zero, thus the uncertainties from BayesSHAP goes to zero as well. On the other hand, GP-SHAP uncertainties still remain valid because it represents the GP predictive uncertainties, which do not shrink to zero as we increase the number of coalitions we use to esitmate the SVs.

Further illustrations: In Figure 5, we plot two violin plots to illustrate the relationship between mean and standard deviation of the stochastic values with respect to the size of the original feature. We see that the feature “worst fractal dimension” are the second most influential feature in terms of mean stochastic explanations and also the feature that has highest uncertainty around its explanations. In comparison with the housing prediction problem illustrated in Figure 4, the higher the feature value doesn’t necessary give higher uncertainty around its explanation.

C.3 Predictive explanations

For this illustration, we utilise the Diabetes dataset [47] with 442 patient data and 10 numeric features measuring the following:

- age: age in years
- sex
- bmi: body mass index
- bp: average blood presuure
- s1: total serum cholesterol
- s2: low-density lipoproteins
- s3: high-density lipoproteins
- s4: total cholesterol
- s5: Log of serum triglycerides level

- s6: blood sugar level

The experiment is to assess the effectiveness of the Shapley prior we proposed in predicting explanations estimated using SHAP algorithms for general models, including GP-SHAP, TreeSHAP, and DeepSHAP. We use the implementation of TreeSHAP and DeepSHAP from the **shap** package [2].

While algorithms such FastSHAP [22] also learn a vector-valued function that returns explanations given instances, the algorithm require access to the underlying model f during training while ours required previously computed explanations. Due to this importance difference in the problem setup, we do not compare the two algorithm.

We first generate three sets of explanations to set up three regression problems:

1. Fit a Gaussian process model and then run GP-SHAP to obtain explanations.
2. Fit a random forest model and then run TreeSHAP to obtain explanations.
3. Fit a neural network model and then run DeepSHAP to obtain explanations.

After obtaining explanations as groundtruths for this experiment, we randomly divide 70% of them as training data and 30% of them as testing data. We then do the following,

1. We fit a multi-output GP using the proposed Shapley prior on the training data and predict the explanations of the unseen test data.
2. We fit a multi-output random forest model on the training data and predict the explanations of the unseen test data.
3. We fit a multi-output neural network model on the training data and predict the explanations of the unseen test data.

We repeat this experiment 10 times using different seeds and compute the RMSE between the predicted and groundtruths explanations. The results are then plotted in Figure 3.

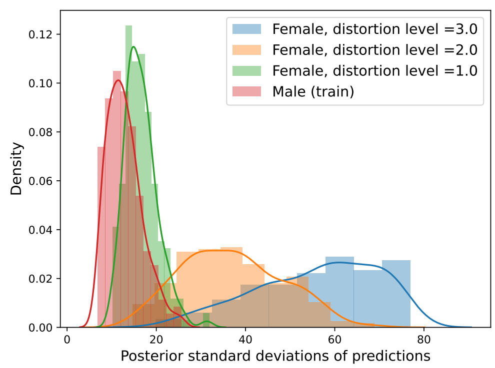
C.4 Further ablation study: Impact of increased posterior prediction uncertainty on explanation uncertainties

In this ablation study, we aim to examine the effect of increasing the uncertainty in posterior predictions on the corresponding uncertainty in stochastic Shapley values. To demonstrate this, we utilize the diabetic dataset [47] and split the data based on recorded sex. We train our GP model on the male data and employ BayesGP-SHAP to explain the prediction results for both the male training data and the female testing data. We adopt this split because we expect the biological characteristics between males and females to be distinct enough to treat the female data as out-of-sample data, thereby naturally resulting in increased predictive uncertainty for the female data. To further amplify this uncertainty, we multiply each instance in the female testing data by distortion factors of two and three, respectively, and assess the corresponding uncertainties in the explanations.

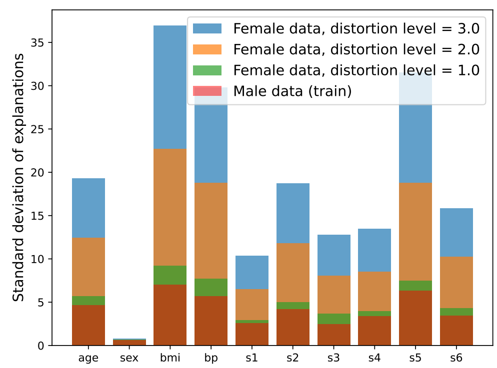
We begin by illustrating the relationship between the out-of-sampleness of the data and the corresponding increase in predictive posterior uncertainties. This is depicted in Figure 6a, where we observe that as the data becomes more out-of-sample, the predictive uncertainties consistently rise. Even at distortion level 1, which represents the original female data, we can already observe increased uncertainties compared to the uncertainties derived from male data prediction.

Furthermore, these increased uncertainties in the predictive posterior are reflected in the associated feature explanations. This is evident in Figure 6b, where we visualize the uncertainties associated with the feature explanations. For instance, the green bars representing the average uncertainties in explaining female data with no distortion are consistently larger than the red bars, which represent the average uncertainties of male data explanations. This observation aligns with the higher predictive uncertainties observed in Figure 6a for the female data compared to the male training data.

It is worth noting that the uncertainty for the feature “sex” remains consistently close to zero. This is because the feature “sex” is constant within both the female and male datasets. As a result, it acts as the null player in each dataset and obtains an almost Dirac zero as its stochastic Shapley value.



(a) Predictive posterior standard deviation



(b) Mean of standard deviations of explanations

Figure 6: Ablation study: (left) We begin by training a Gaussian Process (GP) model on the male data. We then make predictions using this trained model on both the male data and out-of-sample female data. To assess the impact of increasing posterior uncertainties, we multiply the female data by distortion levels of 1.0, 2.0, and 3.0. We visualize the results by plotting the density plot of the standard deviations obtained from the predictive posterior distributions. (right) Next, we focus on analyzing the average standard deviations of explanations per feature from the male and female data, considering different distortion levels. We observe that as we progressively increase the posterior uncertainties in the sample, these uncertainties are reflected in the uncertainties of the explanations provided.