Boundary-Aware Uncertainty for Feature Attribution Explainers

Davin Hill, Aria Masoomi, Max Torop, Sandesh Ghimire, Jennifer Dy

Department of Electrical and Computer Engineering, Northeastern University, Boston, MA

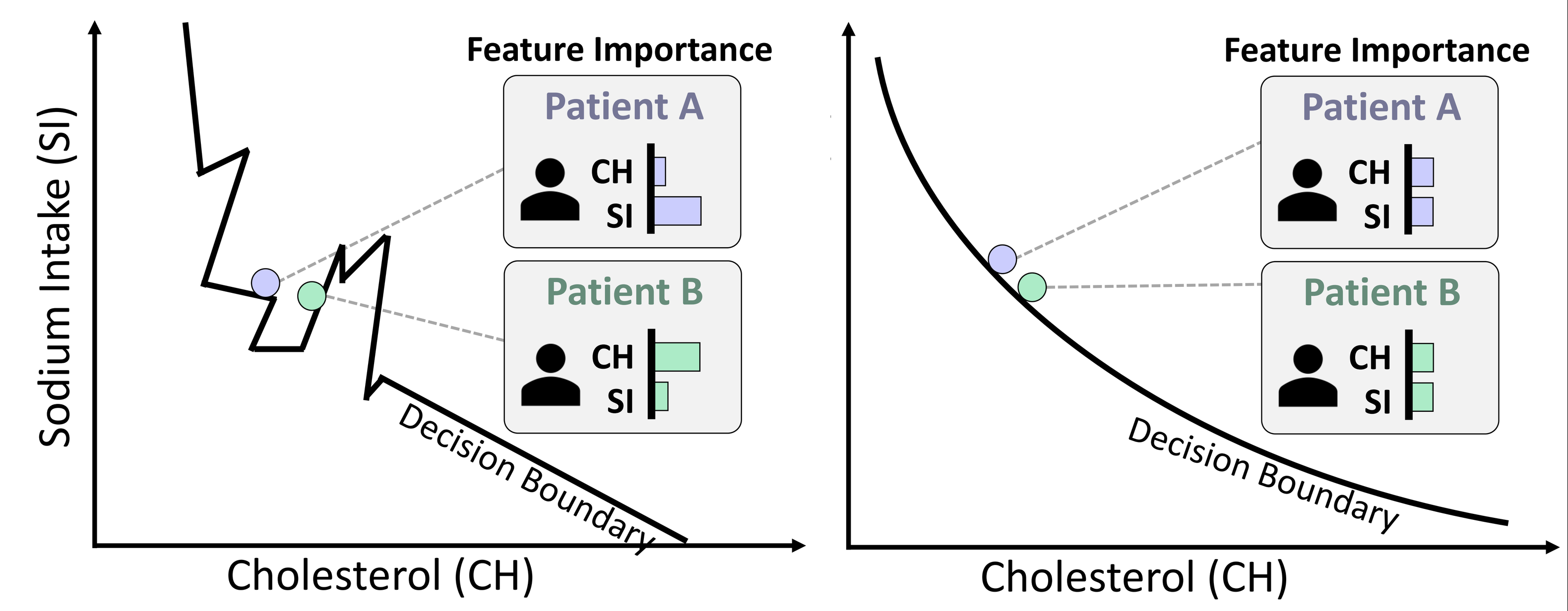
Northeastern University

TL; DR

- We introduce the **GPEC** framework to quantify uncertainty for feature attribution explanations.
- We propose the **Weighted Exponential Geodesic (WEG) kernel** which captures local decision boundary complexity between samples.
- GPEC uses a Gaussian process model to combine uncertainty due to **1) local decision boundary complexity** and **2) function approximation**.

Motivation

- Local Feature Attribution Explainers generate importance scores for any given sample and prediction model.
- Explainers are function approximators of the prediction model; this approximation can have estimation uncertainty.
- Explainers can also be inconsistent, and/or unstable, especially for samples near nonlinear/complex decision boundaries. Therefore these regions may engender higher uncertainty
- Given two samples and their explanations, how much information does one explanation provide about the other?



GPEC Framework

Idea: Model Explainers as latent functions with noisy "labels" (i.e. their explanations E_n)

$$E_{n,d} = \underbrace{\mathcal{H}_d(X_n)}_{\text{Decision Boundary – Aware Uncertainty}} + \underbrace{\eta_{n,d}}_{\text{Function Approximation Uncertainty}}$$

Decision Boundary – Aware Uncertainty

$$\mathcal{H}_d(X_n) \sim \mathcal{GP}(0, k(\cdot,\cdot))$$

- The distribution over *functions* that may have generated the observed explanations
- Kernel $k(\cdot,\cdot)$ encodes similarity between samples; the novel WEG kernel can capture local decision boundary complexity.

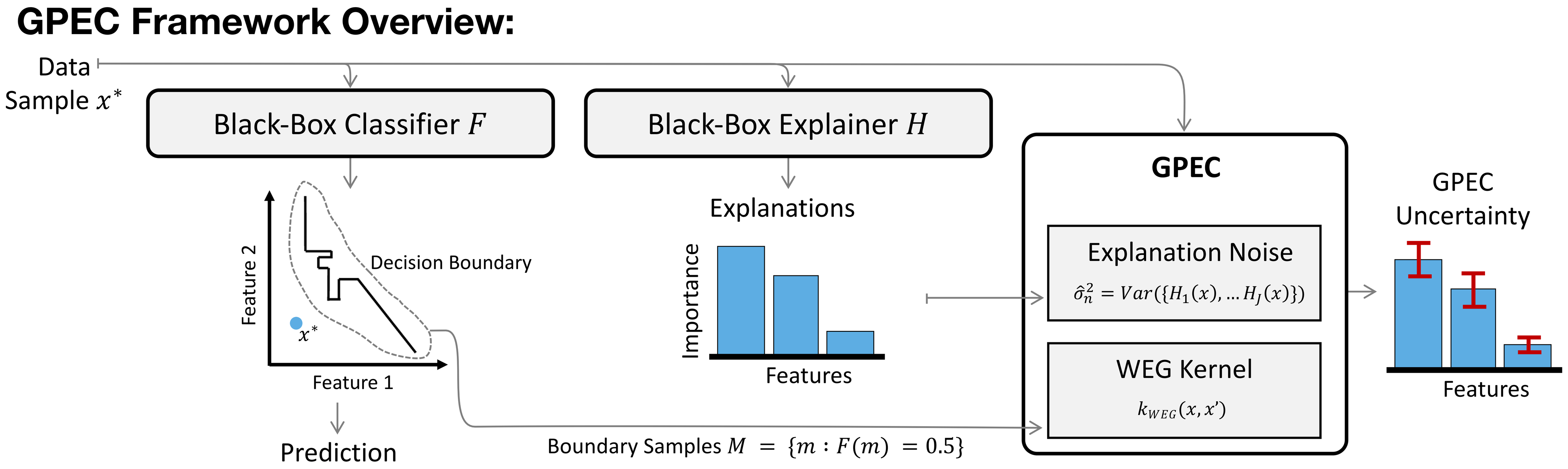
Function Approximation Uncertainty

$$\eta_{n,d} \sim \mathcal{N}(0, \sigma^2_{n,d})$$

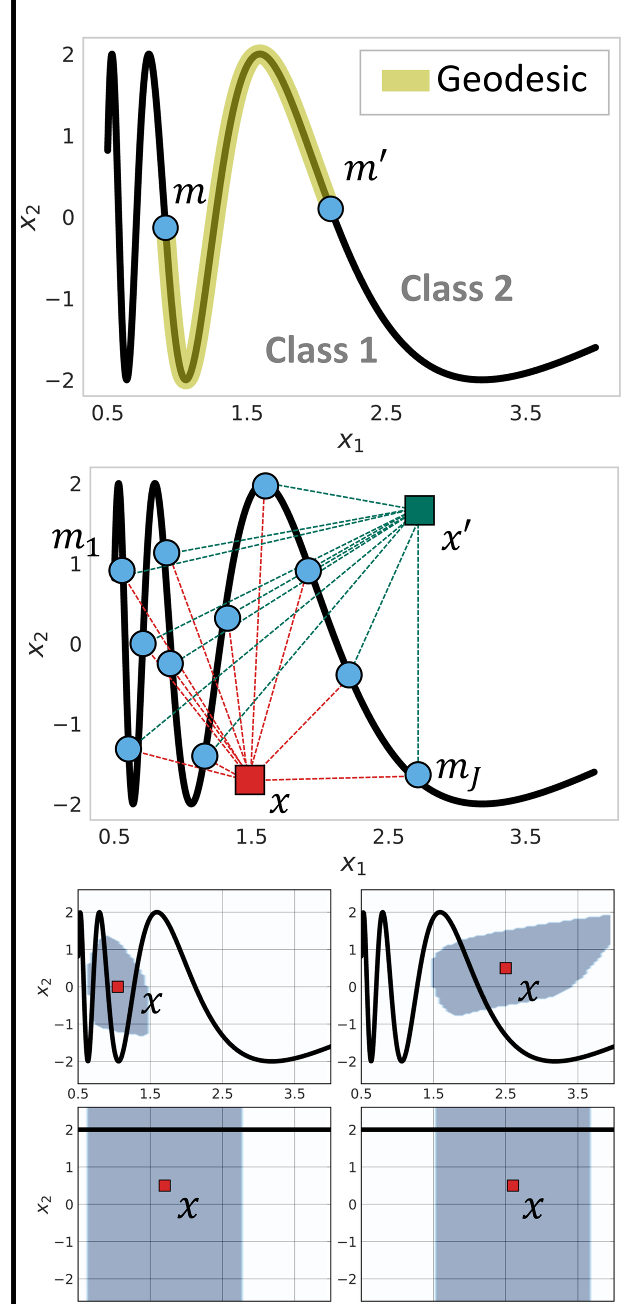
- Uncertainty from stochastic explainers
- Can be used with e.g. BayesLIME / BayesSHAP [2] or estimated empirically.

GPEC Combined Uncertainty for a given sample: Variance of the GP Predictive Distribution $\mathbb{V}_d[x^*] = k(x^*, x^*) - k(X, x^*)^\intercal [K + \sigma^2_d I_N]^{-1} k(X, x^*)$

GPEC Framework Overview:



Weighted Exponential Geodesic Kernel



How can decision boundary complexity be quantified?

⇒ Geodesic distance between points m, m'

Exponential Geodesic (EG) Kernel [1]

$$k_{\mathrm{EG}}(m, m') = \exp\left[-\lambda d_{geo}(m, m')\right]$$

We extend the EG kernel **by weighting each boundary sample by distance to data samples**:

$$q(m|x, \rho) \propto \exp\left[-\rho ||x - m||_2^2\right] p(m)$$

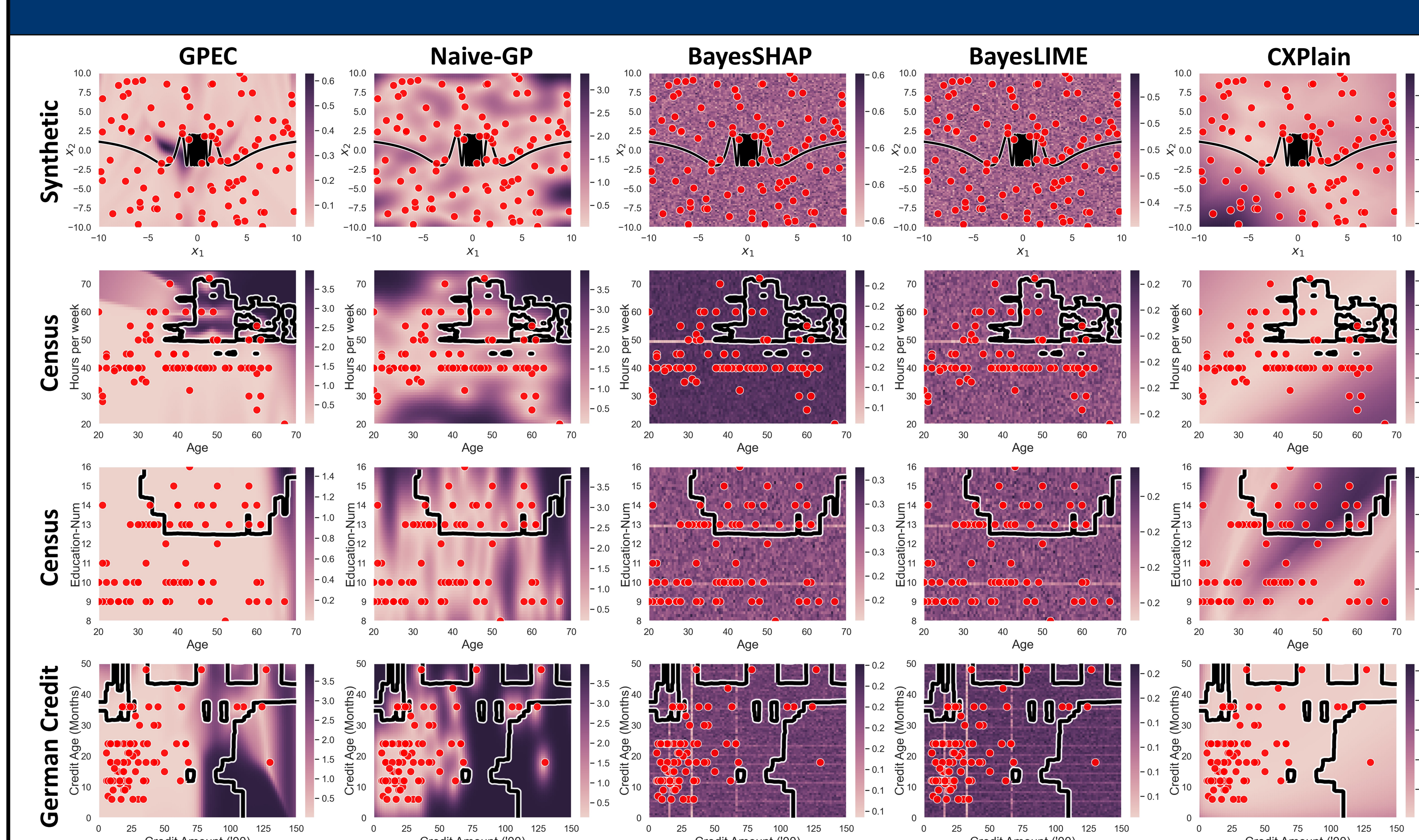
Weighted Exponential Geodesic (WEG) Kernel

$$k_{\mathrm{WEG}}(x, x') = \int\int k_{\mathrm{EG}}(m, m') \times q(m|x, \rho)\, q(m'|x', \rho)\, \mathrm{d}m\mathrm{d}m'$$

Theorem 1. *Given two points $x, x' \in \mathcal{X} \cap \mathcal{M}_F$, then* $\lim_{\rho\to\infty} k_{WEG}(x, x') = k_{EG}(x, x')$

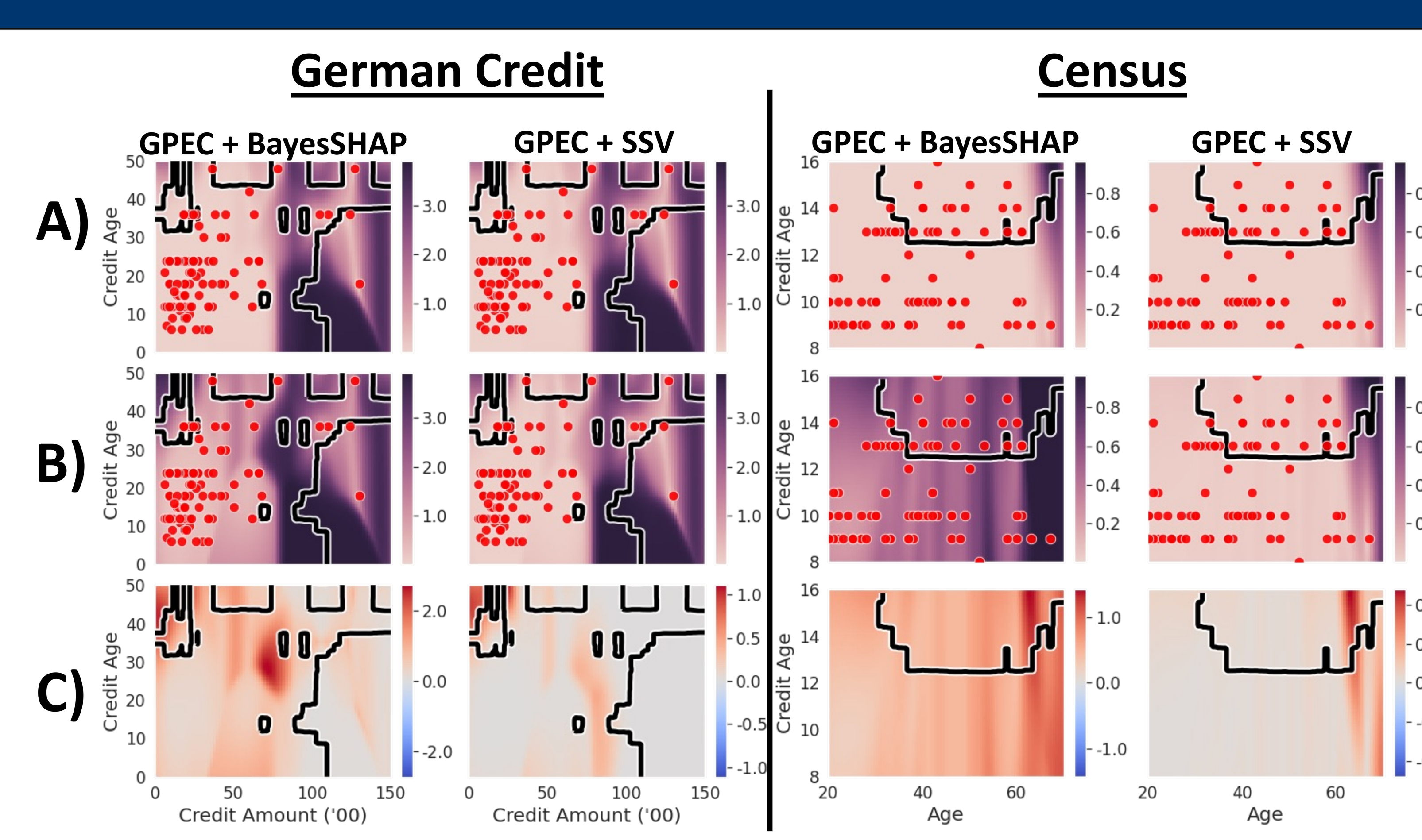
Theorem 2. *Let $\mathcal{P}$ be a $(d-1)$-dimension Piecewise Linear manifold embedded in $\mathbb{R}^D$. Let $\widetilde{\mathcal{P}}$ be a perturbation of $\mathcal{P}$ and define $\tilde{k}(x, x')$ and $k(x, x')$ as the WEG kernels defined on $\widetilde{\mathcal{P}}$ and $\mathcal{P}$ respectively. Then* $\tilde{k}(x, x') < k(x, x')\ \forall x, x' \in \mathbb{R}^D$.

Experiments



Left: Heatmap of Uncertainty values (**Darker = Higher Uncertainty**). GPEC Uncertainty is higher around DB nonlinearities.

Right: Ablation test of GPEC Uncertainty. **A)** full GPEC uncertainty. **B)** DB-aware uncertainty. C) Difference between A and B.



Dataset	Census			Online Shoppers			German Credit			CIFAR10					
Regularization	γ			γ			γ			ℓ_2			Softplus β		
Magnitude	0	5	10	0	5	10	0	5	10	0	1e-5	10e-5	1.0	0.5	0.25
GPEC	**1.573**	**1.177**	**1.158**	**0.209**	**0.123**	**0.092**	**2.665**	**1.747**	**0.250**	**0.029**	**0.029**	**0.028**	**0.033**	**0.032**	**0.032**
Naive-GP	**0.498**	**0.472**	**0.467**	3.699	3.724	3.730	0.330	0.412	2.533	0.902	0.902	0.902	0.903	0.903	0.903
BayesSHAP	0.037	0.037	0.037	0.031	0.031	0.031	0.019	0.019	0.018	–	–	–	–	–	–
BayesLIME	**0.097**	**0.096**	**0.095**	**0.098**	**0.097**	**0.093**	**0.085**	**0.066**	**0.045**	–	–	–	–	–	–
CXPlain	0.064	0.064	0.069	0.004	0.003	0.006	1.7e-4	1.1e-4	4.3e-4	9.5e-5	0.2e-5	2.6e-5	1.6e-5	0.2e-5	9.1e-5

Left: Change in uncertainty as model is trained with increasing regularization.

Acknowledgements / References

[1] Feragen et al. *Geodesic exponential kernels: When curvature and linearity conflict.*

This work was supported in part by NIH 2T32HL007427-41 and U01HL089856 from the National Heart, Lung, and Blood Institute; NIH R01CA240771 and U24CA264369 from the National Cancer Institute; and the Institute of Experiential AI at Northeastern University.