Paper 169: Level Set Estimation with Search Space Warping

Manisha Senadeera, Santu Rana, Sunil Gupta, and Svetha Venkatesh

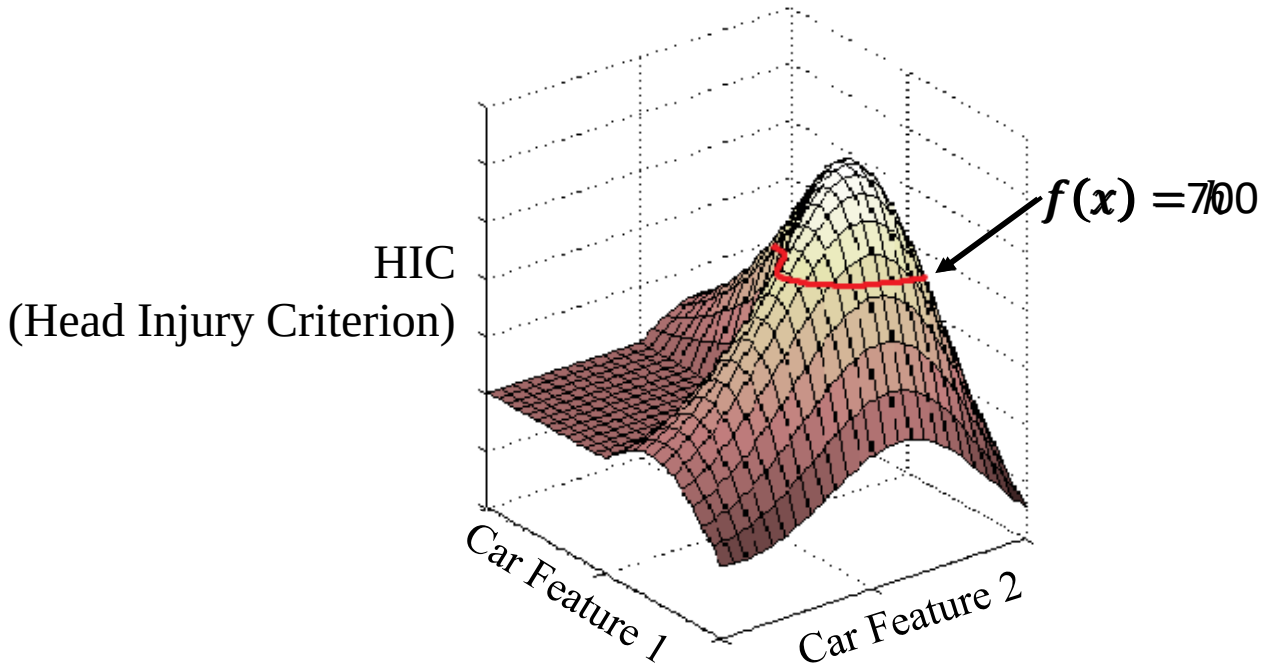
Applied Artificial Intelligence Institute (A2I2), Deakin University, Geelong, Australia
{manisha.senadeera,santu.rana,sunil.gupta,svetha.venkatesh}@deakin.edu.au

A2I2
APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

Motivation

Level Set Estimation - Discover set of ‘designs’ that meet a target

E.g. industrial design – car crash safety



Problem Definition

Find the level set of a function:

$$D_h = \{\mathbf{x} : f(\mathbf{x}) = h\}$$

Where

$\mathbf{x} \in \chi \subset \mathbb{R}^D$

h : desired level

η : small tolerance

$H = \{\mathbf{x} : f(\mathbf{x}) > h\}$: super-level set

$L = \{\mathbf{x} : f(\mathbf{x}) < h\}$: sub-level set

Background

Bayesian Optimisation (BO)

- Global optimisation of expensive black-box functions
- Builds a probabilitic model of the function – using a Gaussian process (GP) prior.
- Posterior used to construct acquisition function, which is optimised to select next sample point

Bayesian Optimisation for Level Set Estimation

Active Learning for Level Set Estimation:

Gotovos, A., Casati, N., Hitz, G., Krause, A.: Active learning for level set estimation. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. pp. 1344–1350. IJCAI '13, AAAI Press (2013)

Minimise distance between mean and desired level

Maximise Uncertainty

$$a_t(\mathbf{x}) = - \mid \mu_{t-1}(\mathbf{x}) - h \mid + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x})$$

$$\mathbf{x}_t = \underset{x}{\operatorname{argmax}}\, a_t(\mathbf{x})$$

Proposed solution

Complex covariance function with a kernel to warp (expand and contract) the search space used by the acquisition function to select the next sample point

Expand regions where level has a high chance of existing

Contract regions where chance is low

How?

Snoek, J., Swersky, K., Zemel, R., Adams, R.P.: Input warping for bayesian optimization of non-stationary functions. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. pp. II–1674–II–1682. ICML'14, JMLR.org (2014)

By constructing a complex covariance function via a **non-homogenous length scale**.

The form of the kernel is:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \mid \Sigma_i \mid^{\frac{1}{4}} \mid \Sigma_j \mid^{\frac{1}{4}} \mid \frac{\Sigma_i + \Sigma_j}{2} \mid^{-\frac{1}{2}} g(\mathbf{x}_i, \mathbf{x}_j)$$

$$g(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-(\mathbf{x}_i - \mathbf{x}_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2} \right)^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right]$$

$$\Sigma_i = \mathrm{diag}(\boldsymbol{l}(\mathbf{x}_i)^2)$$

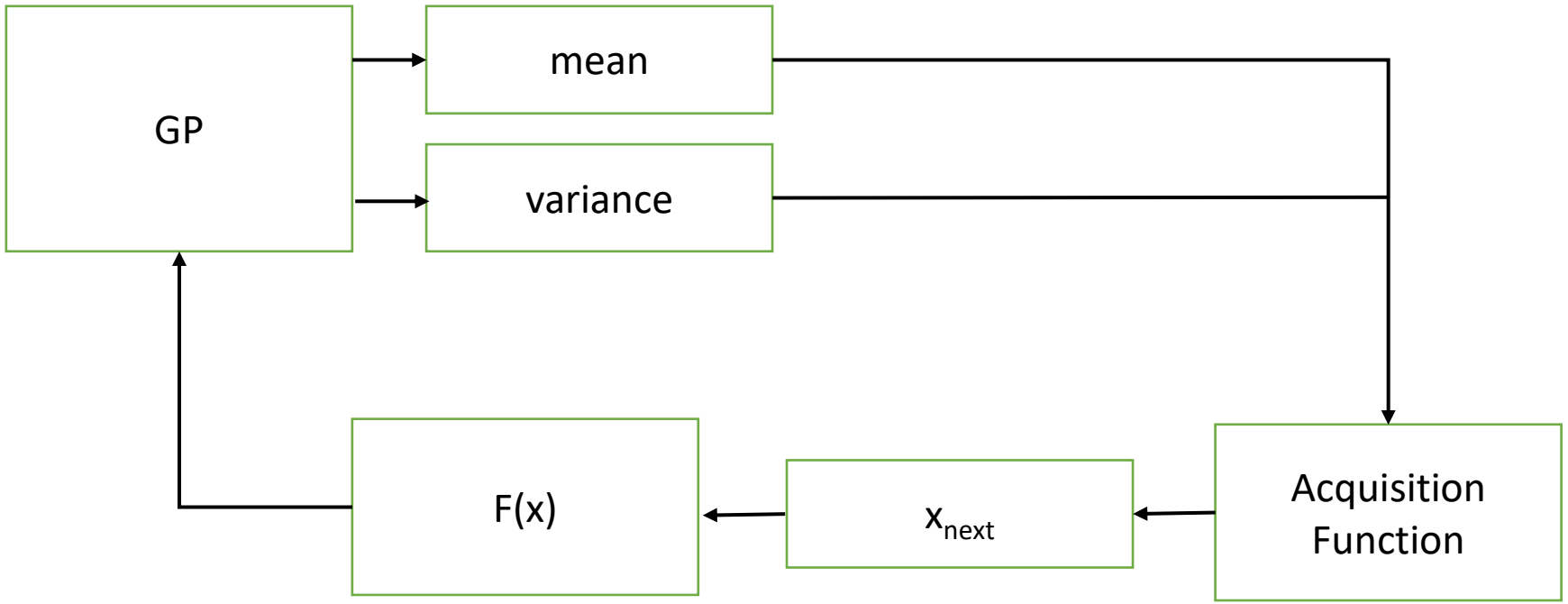
$\boldsymbol{l}(\mathbf{x}_i)$: vector of length scales

Warping Length Scale

Function to determine the new length scale:

$$\boldsymbol{l}(\mathbf{x}) = \boldsymbol{l}_0 \log\left(1 + \left(\frac{\mid \mu_{t-1}(\mathbf{x}) - h \mid + \epsilon}{\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}) + \epsilon}\right)^2\right) + \boldsymbol{l}_1$$

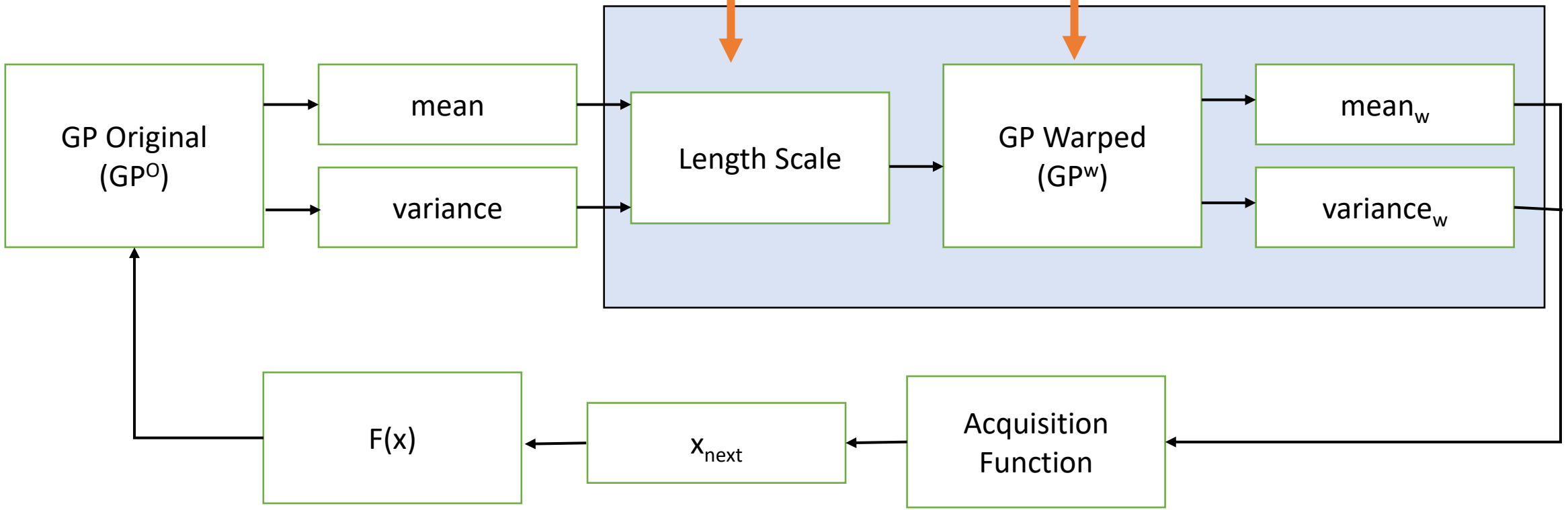
Standard Algorithm



Proposed Algorithm

$$l(\mathbf{x}) = l_0 \log\left(1 + \left(\frac{|\,\mu_{t-1}(\mathbf{x}) - h\,| + \epsilon}{\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}) + \epsilon}\right)^2\right) + l_1$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \mid \Sigma_i \mid^{\frac{1}{4}} \mid \Sigma_j \mid^{\frac{1}{4}} \mid \frac{\Sigma_i + \Sigma_j}{2} \mid^{-\frac{1}{2}} g(\mathbf{x}_i, \mathbf{x}_j)$$



Theoretical Guarantees

Theorem 1: Provides bounds on the length scale of the complex covariance function

Theorem 1. *For any* $h \in \mathbb{R}$, *let* $\delta \in (0,1)$ *and* $\beta_t = 2 \parallel f \parallel_k^2 + 300\gamma_t \, log(t/\delta)^3$, *then with probability* $\geq 1-\delta$, *the length scale will be bounded between* $\boldsymbol{l}_1 \leq \boldsymbol{l}(\boldsymbol{x}) \leq \boldsymbol{l}_0$ $log\left(1+\left(1+\frac{\Delta_{fmax}}{\epsilon}\right)^2\right)+\boldsymbol{l}_1$, *where* $\Delta_{fmax} = max| \, f(\boldsymbol{x}) - h \, |$

Theorem 2: Convergence of cumulative regret using acquisition function (with true length scale) in a continuous domain.

Theorem 2. *Let* $\delta \in (0,1)$, $\beta_t = 2 \parallel f \parallel_k^2 + 300\gamma_t^w ln^3(t/\delta)$, γ_t^w *be maximum information gain for the warped squared exponential kernel after t iterations,* σ^2 *be variance of the measurement noise and h be the desired threshold. Then with probability of* $\geq 1-2\delta$, *the cumulative regret of the ambiguity acquisition function of (3) follows the sublinear rate* $R_T \leq \sqrt{\frac{8T\beta_T\gamma_t}{log(1+\sigma^{-2})}} + T \mid f(\boldsymbol{x}^*) - h \mid$.

Theorem 3: Maximum information gain is bounded even under the heterogeneous length scale range described in Theorem 1

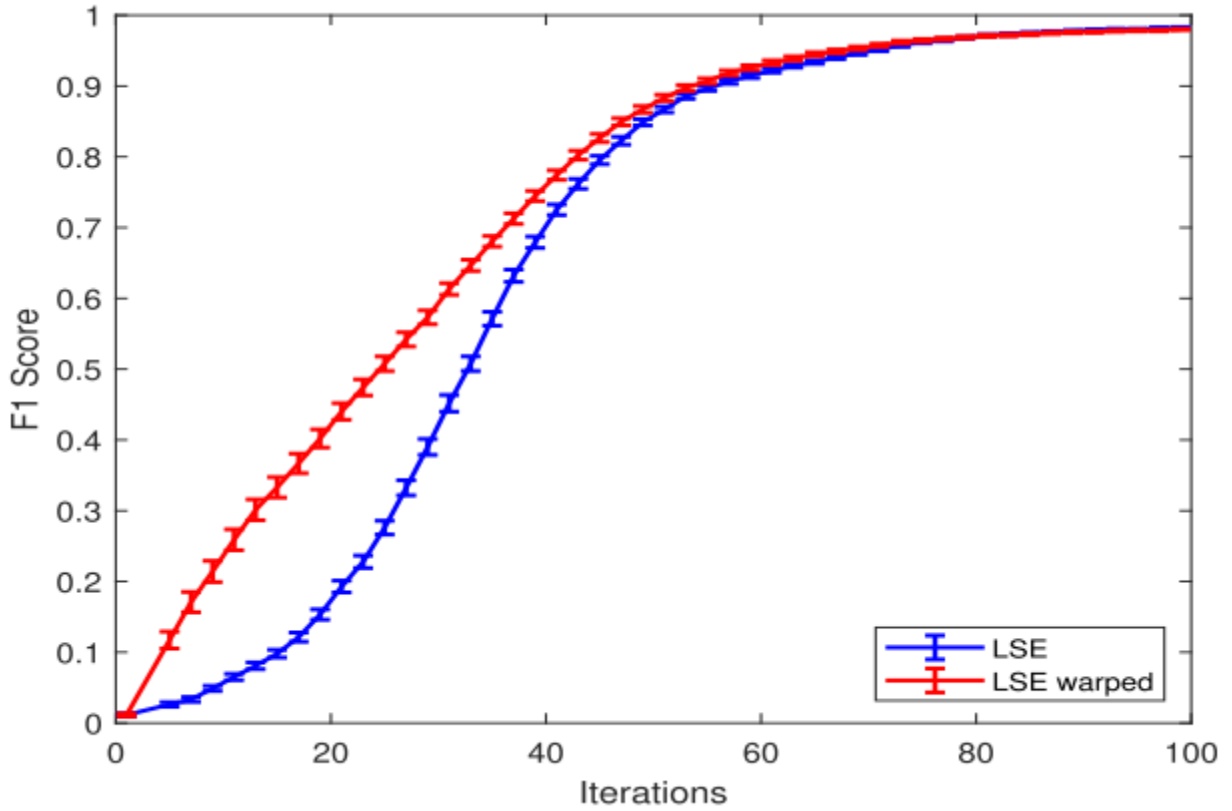
Theorem 3. *Let* $D \subset \mathbb{R}^d$ *be compact and convex,* $d \in \mathbb{N}$. *Assume the kernel function satisfies* $k(\boldsymbol{x}, \boldsymbol{x}') \leq 1$. *Then for our proposed covariance function with varying length scale as described in (6), the maximum information gain at iteration* T *is* $\mathcal{O}((\log T)^{d+1})$.

Experimental Results

Mishra's Bird Function

$$f(x,y) = sin(x)e^{(1-cos(y))^2} + cos(y)e^{(1-sin(x))^2} + (x-y)^2$$

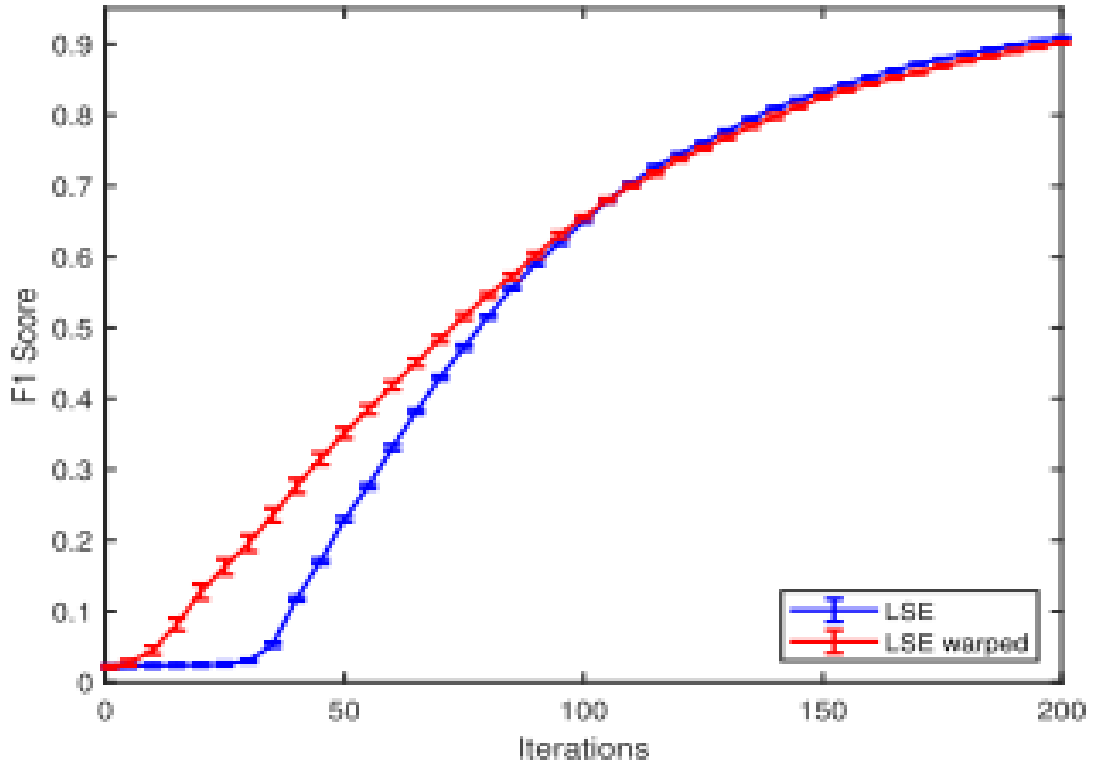
$$h = 10$$



F1 vs iteration results for classification of Bird Function at level h = 10

Car Crashworthiness Design

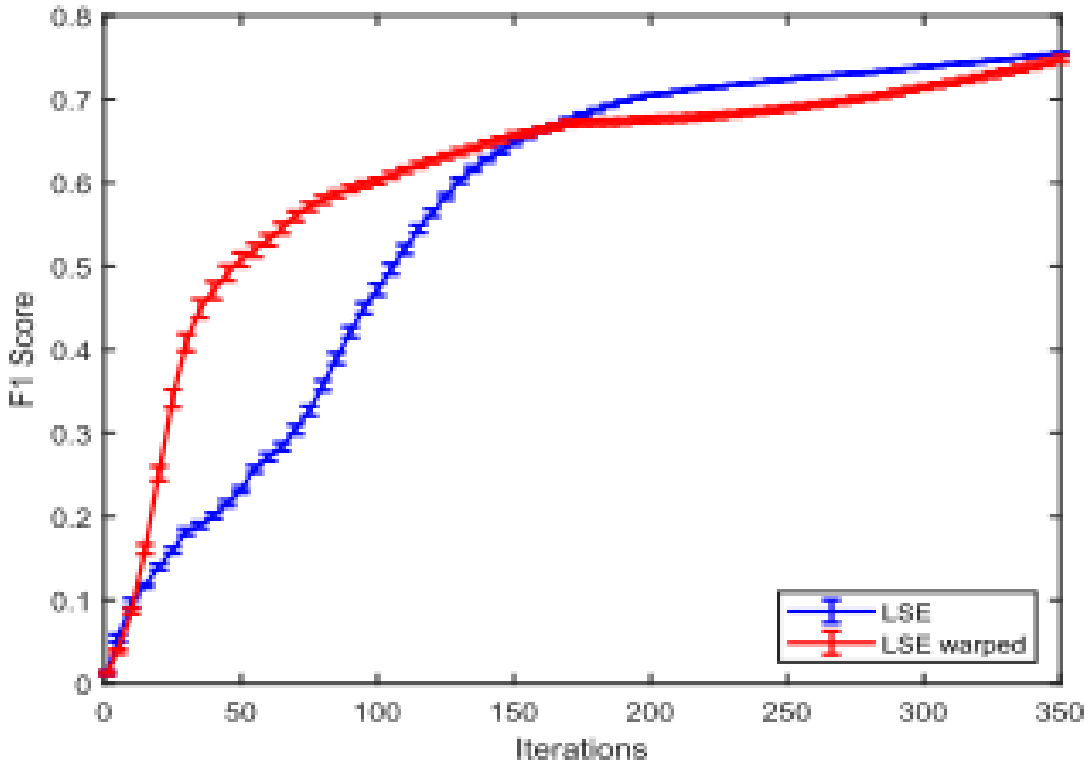
$h = 250$



(a) Experiment 1 results

F1 score vs iteration for HIC < 250

$h = 1.9\ Hz$

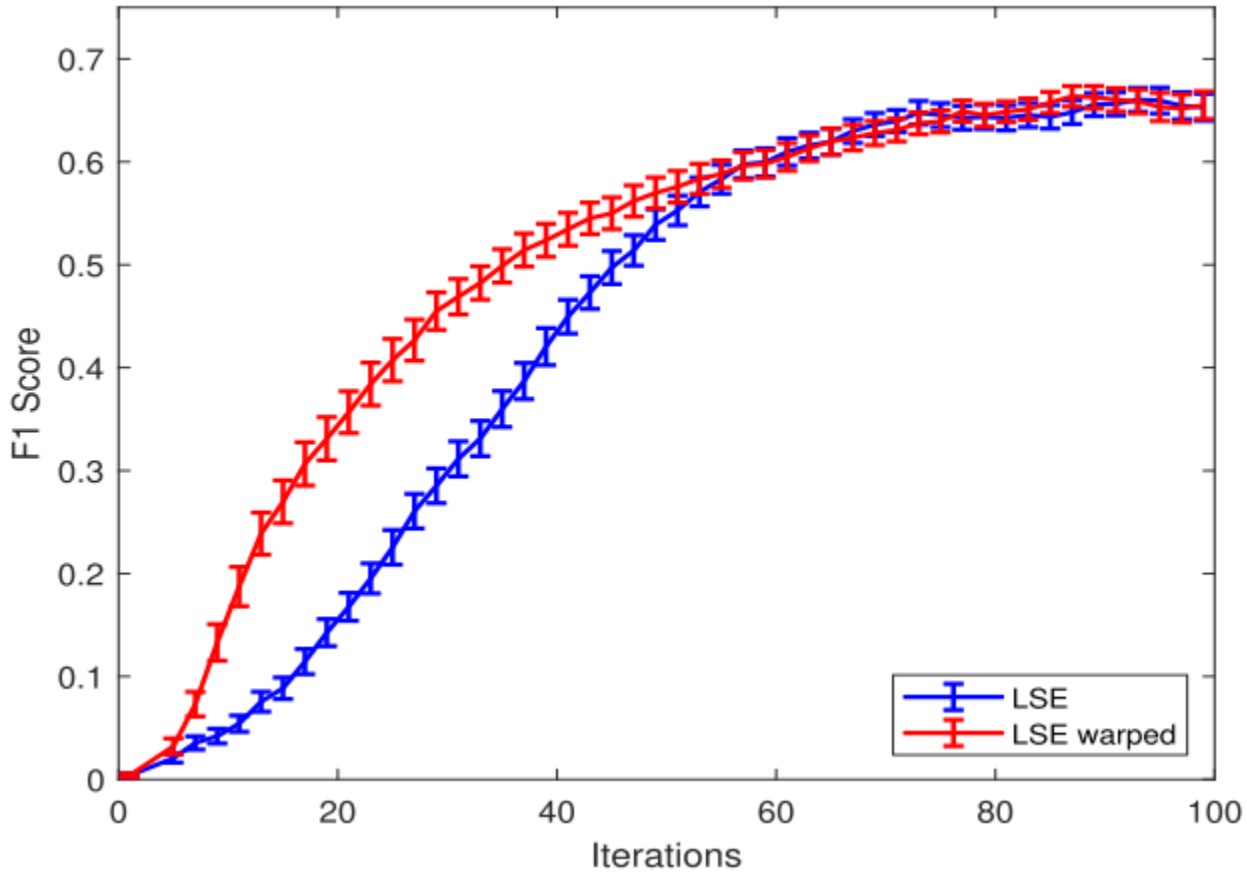


(b) Experiment 2 results

F1 score vs iteration for torsional mode frequency < 1.9Hz

Ductile Alloy Design

$$h = 0.8$$



F1 vs iteration results for classification of alloy with level at 80% FCC

Computational Time

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{d,i} - x_{d,j})^2}{l_d^2}\right)$$

50 iterations ≈ 12 seconds

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \mid \Sigma_i \mid^{\frac{1}{4}} \mid \Sigma_j \mid^{\frac{1}{4}} \mid \frac{\Sigma_i + \Sigma_j}{2} \mid^{-\frac{1}{2}} g(\mathbf{x}_i, \mathbf{x}_j)$$

$$g(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-(\mathbf{x}_i - \mathbf{x}_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mathbf{x}_i - \mathbf{x}_j)\right]$$

50 iterations ≈ 5 minutes

Conclusion

- Presented a novel means to construct a complex covariance function by distorting the length scale of the GP from which the acquisition function samples from
- Method expands and contracts area to encourage more sampling in regions with high potential for being at the desired level.
- Method results in acquisition function operating more exploitatively.
- Convergence guarantees are presented.

Thank you