Level Set Estimation with Search Space Warping

Manisha Senadeera(✉), Santu Rana, Sunil Gupta, and Svetha Venkatesh

Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Geelong, Australia
{manisha.senadeera,santu.rana,sunil.gupta,
svetha.venkatesh}@deakin.edu.au

Abstract. This paper proposes a new method of level set estimation through search space warping using Bayesian optimisation. Instead of a single solution, a level set offers a range of solutions each meeting the goal and thus provides useful knowledge in tolerance for industrial product design. The proposed warping scheme increases performance of existing level set estimation algorithms - in particular the ambiguity acquisition function. This is done by constructing a complex covariance function to warp the Gaussian Process. The covariance function is designed to expand regions deemed to have a high potential for being at the desired level whilst contracting others. Subsequently, Bayesian optimisation using this covariance function ensures that the level set is sampled more thoroughly. Experimental results demonstrate increased efficiency of level set discovery using the warping scheme. Theoretical analysis concerning warping the covariance function, maximum information gain and bounds on the cumulative regret are provided.

Keywords: Level set estimation · Gaussian processes · Bayesian optimisation

1 Introduction

Level set estimation is a common problem in industrial design where, instead of a single best design, it is useful to find a set of designs that meet a target. This can then be used for robust manufacturing or to further screen on subsidiary objectives. Consider designing the structure of a vehicle to achieve target crash-safety performance. Vehicle regulations require that the Head Injury Criterion (HIC) not exceed 700 under standard test conditions [6], with lower values indicating better protection against brain injury. The design process often involves first generating a set of designs that meet the criteria via a crash simulator [9] and then filtering for cost before selecting one for actual ground testing. Generating the set of designs (or a representative set, in the case of continuous variables) can be posed as a level set estimation problem. Coupled with the fact that such simulators are computationally expensive, it is important that level sets are found in the minimum number of trials. Similarly, in alloy design, it is useful to find the set of elemental composition that produce alloys with similar mechanical properties. Such a set can then be used for robust specification of the alloy composition. Similar examples are abound in other domains [7]. Hence, level set estimation is an important problem.

Bayesian optimisation (BO) is a method for global optimisation of expensive black-box functions [3]. It has been adapted to seek a level set instead of the optimum [5–8].

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12085, pp. 827–839, 2020.
https://doi.org/10.1007/978-3-030-47436-2_62

It works by building a probabilistic model of the function (normally using a Gaussian process (GP) prior) and then using the posterior to seek the next sample point such that more samples are obtained from the level set. The search for the next point is guided by the optimisation of a surrogate model, known as an acquisition function. There are two major limitations in the current work on BO for level set estimation: 1) they considered only discrete sets to demonstrate convergence of the algorithm, and 2) they end up being more explorative i.e. samples are more scattered as they do not use the fact that most of the level sets for a continuous function tend to be contiguous. The first limitation restricts its application and the second provides an avenue to further improve the sample efficiency, particularly under a budget constraint. Hence, scope for a level set estimation algorithm for expensive black-box functions that works with continuous variables and exploits the continuity of the functions for improved efficiency is still open.

We present a framework of BO for level set estimation using a warped GP to exploit the continuity of the function and then analyse its convergence both for continuous and discrete cases. We implement the warped GP through a non-stationary covariance function such that regions with high potential to be on the level are expanded, whilst others are contracted. The potential is computed by a monotonic function of the difference of the mean prediction from the intended level, scaled by the predicted variance (both the predictions are obtained from an intermediate GP with stationary kernel). The difference from mean term encourages areas close by to existing samples from the level set to have high potential (using continuity of the GP), whilst the variance scaling guards against any undue optimism. When compared against the usual stationary kernel, it tends to operate in a more exploitative manner. Under a budget constraint, exploitative sampling is more beneficial as once one point at the level is discovered, sampling close by is likely to reveal more points on that level. In contrast an explorative algorithm may find little or no points on the level before the budget expires. Theoretical analysis of our proposed warped kernel based approach shows the proposed algorithm is able to retain the sublinear growth rate of the cumulative regret and extensive experiments with both synthetic and real-world functions (including on alloy and car design) demonstrate a significant increase in sample efficiency.

1.1 Related Work

Previous work into level-set estimation problems have been performed by the LSE algorithm [8] and the Truncated Variance Reduction (TruVar) algorithm [2]. In LSE, the ambiguity acquisition function is adapted from the Straddle heuristic [4], providing a balance between exploration and exploitation. This algorithm is an online method that utilises confidence intervals to classify points as being either above or below the level. Similar in nature to LSE in its classification, the TruVar algorithm provides functionality for both Level set estimation and BO applications, utilising the common GP based approach to unite the methods. For Level set estimation applications, the algorithm uses lookahead to select the next sample point as one which provides the greatest reduction in the sum of truncated variances within a set of unclassified points. This method further incorporates point-wise costs and heteroscedastic noise into its selection. In the application of this method in both [8] and [2], the authors have utilised a monotonically decreasing set for unclassified points, based on the bounds of the GP. This method of classification however is suited when sampling is done on a discrete domain.

For super-level set estimation, [16] proposes Maximum Improvement for Level-Set Estimation (MILE), a one-step lookahead algorithm, to locate points that exceed a threshold with a specified high probability. Aiming to find the largest region that exists above a certain level, it operates by sampling points which provide the greatest expected improvement in the set of points classified as being above the threshold. Convergence guarantees were provided even for misspecified prior distribution, however, only for discrete domains.

Level set estimation lends itself to estimating system probability of failure. Work by [1] develops a Bayesian framework for such tasks. They propose a one step-look ahead sequential sampling strategy called stepwise uncertainty reduction (SUR). In [5] this method is adapted from sequential to batch sampling, allowing for parallel sampling of the function. Convergence analysis was not provided.

1.2 Problem Definition

We assume a function $f : \mathbf{x} \rightarrow y$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^D$ is a compact subset from a D-dimensional real vector space and $y \in \mathbb{R}$ is from the real line. We wish to find the level set of a function i.e.

$$D_h = \{\mathbf{x} : f(\mathbf{x}) = h\} \tag{1}$$

where h is the desired level. A small tolerance η is permitted. For some problems it may be useful to find a super-level set i.e. $H = \{\mathbf{x} : f(\mathbf{x}) > h\}$ or a sub-level set i.e. $L = \{\mathbf{x} : f(\mathbf{x}) < h\}$. We assume that the function $f(.)$ is expensive, and only noisy evaluations i.e. $y = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ are available. Hence, we need to find the level set in an sample-efficient way.

1.3 Background

Bayesian optimisation has been adapted for level set estimation of an expensive function, because of its high sample efficiency. Usually a Gaussian process is used to serve the probabilistic model of the function, which is then utilised to select the next sample point using a surrogate function. In the following we outline the Gaussian process and the acquisition function specific to level set estimation.

Gaussian Processes. Gaussian process is a commonly used prior over the space of smooth functions [13]. It is fully defined by a mean and covariance function. Without loss of generality we can assume the mean to be a zero function, then a GP is fully defined by the co-variance function alone, i.e. $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$. Given a set of observations $(\{\mathbf{x}_i, y_i\}_{i=1}^t)$, the posterior is also a GP whose predictive mean and covariance can be computed as, $\mu_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^T(K_t + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}_t$, and $k_t(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_t(\mathbf{x})^T(K_t + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_t(\mathbf{x}')$, with variance $\sigma_t^2(\mathbf{x}) = k_t(\mathbf{x}, \mathbf{x})$, $\mathbf{k}_t(\mathbf{x}) = [k(\mathbf{x}_i, \mathbf{x})]_{i=1}^t$ and, $K_t = [k(\mathbf{x}_t, \mathbf{x}_{t'})]_{t,t'}$ is the kernel Gram matrix.

A popular kernel used for the covariance function is the squared exponential of the form (assuming stationarity):

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

where σ_f^2 is the signal variance, and l_d is a constant length scale for the d-th dimension.

Level Set Estimation Algorithm. We use the algorithm proposed by [8]. Based on the GP model, they used the ambiguity acquisition function $a_t(\mathbf{x})$ to sample the next point. The authors [8] worked on the task of classifying discrete points into super-level and sub-level sets, and the name *'ambiguity'* reflects the uncertainty during the classification process. This acquisition function, described in Eq. (3), aims to minimises the distance between the mean and desired threshold h (exploitation) whilst maximising the uncertainty (exploration).

$$a_t(\mathbf{x}) = - \mid \mu_{t-1}(\mathbf{x}) - h \mid + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}) \tag{3}$$

where β_t trades between exploitation and exploration. The next point is determined as:

$$\mathbf{x}_t = \underset{x}{\operatorname{argmax}}\, a_t(\mathbf{x}) \tag{4}$$

The sequence of β_t can be set in a specific way [15] to achieve an efficient sub-linear convergence rate for cumulative regret ($\triangleq \sum_{i=1}^{t} |f(\mathbf{x_i}) - h|$).

The algorithm can be run either until the iteration budget expires (continuous case) or all the points have been classified between level-set and the rest (discrete case).

2 Problem Setup and Proposed Algorithm

As mentioned, we aim to find the level set of an expensive function with a minimum number of samples. We will make the current process [8] faster by exploiting the contiguous nature of level sets. This is by defining a warping kernel function that expands regions where the level has a higher chance of existing, while contracting regions where the chance is low. This is done by first computing a GP without warping, also referred to as the original GP (GP^o), and using its predictive mean, $\mu_t(\mathbf{x})$ and variance $\sigma_t^2(\mathbf{x})$ to construct the warped kernel used to compute the warped GP (GP^w). GP^w is then used for computing the warped acquisition function. GP^w ensures regions around an observation already at the level is endowed with smaller length-scales than other regions, resulting in higher acquisition function values, thus translating to a higher chance of selecting the next sample from that region. We describe the warping kernel and then analyse its properties. We then provide the warped acquisition function, followed by convergence analysis. We note it is useful to build a better understanding of the level set rather than outputting a small set of sampled points that exist at the level. For such situations we output a GP model based on the samples. As most samples tend to come from near the level, we believe that the level set produced by this GP would be more accurate than when samples come from other means e.g. existing level set estimation methods. This performance can be tested by classifying other points in the region (that are not on the level) into a super-level and sub-level set.

2.1 Input Warping

Snoek et al. [14] warped the input space of non-stationary functions to convert them to stationary functions. We utilise this concept to construct a complex covariance function via a non-homogenous length scale. Because the complex covariance function is

unknown, adapting the length scale instead alleviates the need to pre-define the covariance function. For this, the following form of the kernel [11] is used:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \mid \Sigma_i \mid^{\frac{1}{4}} \mid \Sigma_j \mid^{\frac{1}{4}} \mid \frac{\Sigma_i + \Sigma_j}{2} \mid^{-\frac{1}{2}} g(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

where

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

where Σ_i, known as the kernel matrix, is the covariance matrix of the Gaussian kernel at $\mathbf{x}_i$ [11]. In the isotropic case, this matrix has the form $l_i^2 \times I_D$ [12]. This can be extended to an anisotropic case of the form $\Sigma_i = \text{diag}(\boldsymbol{l}(\mathbf{x}_i)^2)$ where $\boldsymbol{l}(\mathbf{x}_i)$ is a vector of length scales for each dimension at $\mathbf{x}_i$, ensuring $k(\mathbf{x}_i, \mathbf{x}_j)$ remains positive semi-definite.

With the balancing act of the acquisition function being to encourage selection of points which minimise the distance between h and the mean whilst maximising uncertainty, this same objective was incorporated into the length scale warping metric. For problems involving level set estimation an argument can be made for there to be a stronger emphasis on exploitation compared to a BO problem. The reason for this is that unlike BO, once a single point at the desired level is found, it can be safely assumed, for a continuous function, that points close by will also be at that level. Shown in (6) is the metric by which, for a given point, a new length scale value is determined.

$$\boldsymbol{l}(\mathbf{x}) = \boldsymbol{l}_0 \log\left(1 + \left(\frac{\mid \mu_{t-1}(\mathbf{x}) - h \mid + \epsilon}{\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}) + \epsilon} \right)^2 \right) + \boldsymbol{l}_1 \tag{6}$$

The length scale for a point through (6) can be added into the Σ matrix in (5).

By (6), areas with small length scales encourage sampling, as the standard deviation and mean return to prior values of σ_f and 0 faster, behaving like an expanded space. Areas with larger length scales discourage sampling as the mean and standard deviation will remain similar to neighbouring points, mimicking regions with a contraction in size.

A small term, ϵ, in the denominator acts to prevent undefined values, should the uncertainty term reach 0 (as the case for a sampled point without noise). ϵ in the numerator allows uncertainty to still influence warping for points where the mean is equal to the threshold h. Additionally, a 1 is added within the log term to ensure $\boldsymbol{l}(\mathbf{x})$ remains a positive function, and both $\boldsymbol{l}_0$ and $\boldsymbol{l}_1$ are to be positive.

It is necessary that the form of the length scale warping metric be different to that of the acquisition function, whilst still valuing a similar exploration-exploitation balance. This avoids both metrics always preferring the same point (avoiding a doubling up).

It is possible to apply multiple warpings where after the original GP is warped, the resulting GP is again warped multiple times. After this, the acquisition function is applied to the final warped GP to select the next best point. Warping causes the acquisition function to behave more exploitatively, as shown in Fig. 1. Initially the point selected by the acquisition function (indicated by the red square) is more explorative but, after multiple warpings, the point is more exploitative. More exploitation is not necessarily good as samples from the level set would start to look very similar.

The balance lies somewhere in the middle where exploitation is high enough to use the contiguous nature of the level set, but low enough to give variability in the samples. In our experience one level of warping tends to give sufficient exploitation behaviour.

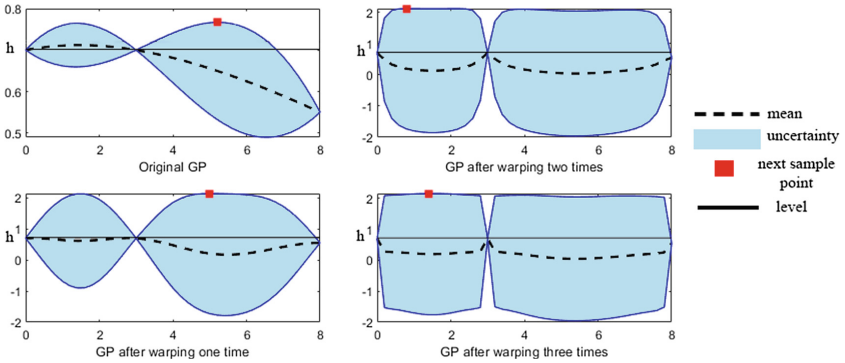
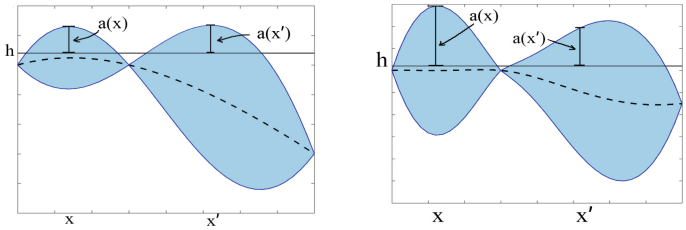


Fig. 1. Impact to selection of next point by acquisition function due to multiple length scale warpings. Increased warping layers result in the acquisition function behaving more exploitatively. (Color figure online)

Figure 2 illustrates the exploitative behaviour of warping compared to the unwarped kernel approach. Without warping, both x and x' have the same acquisition value and are equally likely to be selected. With warping, these points are differentiated as x will have a lower length scale than x', giving it a higher chance of being selected.



(a) Without warping (b) With warping

Fig. 2. (a) Ambiguity acquisition function value equal for both x and x′ - equally likely to be selected as next best point. (b) Differentiation of acquisition value for x and x′ following warping

2.2 LSE with Input Warping

The LSE Algorithm with input warping is described in Algorithm 1. Though the acquisition function itself is unchanged, the search space for selecting the next sample point is warped. As such the form of the acquisition function used is shown in (7).

$$a_t(\mathbf{x}) = -\mid \mu_{w_{t-1}}(\mathbf{x}) - h \mid + \sqrt{\beta_t}\sigma_{w_{t-1}}(\mathbf{x}) \tag{7}$$

where μ_w and σ_w are the mean and variance from a warped GP with the complex covariance function using length scale given by (6). The GP used to classify the test points remains un-warped as it is based on a true model selection approach making it the most accurate model of the function. Classification of points into the super-level H, sub-level L and unclassified U sets follow the same approach as [8], however these sets are no longer monotonic in size and allow for re-classification. Furthermore in [8] sampling for the LSE algorithm was limited to points from the unclassified U set. In Algorithm 1, to allow extension to continuous domain, this constraint was removed.

Algorithm 1. LSE with input warping

Input: Initial data set D_0
Parameter: Desired threshold h. Tolerance η around h.
Output: D_h and GP_T^o

$D_h \leftarrow \emptyset$
for $t = 1, 2, ..., T$ **do**
  Estimate length scale of unwarped space l and, warping length scale parameters $\boldsymbol{l}_0$ and $\boldsymbol{l}_1$ using D_{t-1}
  Compute μ_{t-1} and σ_{t-1} with D_{t-1} and l
  Compute warped length $\boldsymbol{l}(\mathbf{x})$ using (6) and $\boldsymbol{l}_0$ and $\boldsymbol{l}_1$
  Re-fit D_{t-1} to GP_{t-1}^w according to (5) using $\boldsymbol{l}(\mathbf{x})$ and derive $\mu_{w_{t-1}}(\mathbf{x})$ and $\sigma_{w_{t-1}}(\mathbf{x})$
  Choose
$$\mathbf{x}_t = \underset{x}{\operatorname{argmax}} -\mid \mu_{w_{t-1}}(\mathbf{x}) - h \mid + \sqrt{\beta_{t-1}}\sigma_{w_{t-1}}(\mathbf{x})$$
  Sample function $y_t = f(\mathbf{x}_t)$ and update data $D_t = D_{t-1} \cup (\mathbf{x}_t, y_t)$
  Construct GP_t^o from data D_t using l
  if $h - \eta < y_t < h + \eta$ **then**
    $D_h = D_h \cup x_t$
  end if
end for

2.3 Theoretical Guarantees

We provide theoretical guarantees for the proposed method. In Theorem 1 bounds on the length scale for the complex covariance function are provided. Next, we analyse the convergence of the acquisition function, detailed in Theorem 2, with the true length scale by providing a bound on cumulative regret. This bound is described by the maximum information gain. In Theorem 3 we prove that this maximum information gain is bounded even under the heterogeneous length scale range described in Theorem 1. The theorems demonstrate that the convergence rate remains unaffected even with warping. In our empirical analysis, the warped acquisition function performs more efficiently.

Theorem 1 defines bounds on the warped length scale.

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 $\Delta_{fmax} = max\mid f(\boldsymbol{x}) - h \mid$.*

Proof. Proof of Theorem 1 is provided in supplementary materials.

Gotovos et al. [8] provided theoretical convergence bounds for the acquisition function in discrete domain problems by bounding the number of samples required for a specified confidence. Theorem 2 provides a cumulative regret bound for the acquisition function in a continuous domain, where regret is the same as that defined in [15].

Theorem 2. *Let* $\delta \in (0,1)$, $\beta_t = 2 \parallel f \parallel_k^2 + 300\gamma_t^w ln^3(t/\delta), \gamma_t^w$ *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* $R_T \leq \sqrt{\frac{8T\beta_T\gamma_t}{log(1+\sigma^{-2})}} + T \mid f(\boldsymbol{x}^*) - h \mid$.

Proof. Proof of Theorem 2 is provided in supplementary material.

The regret bound assures that the algorithm converges to the desired level, with cumulative regret reflecting the rate of convergence. Theorem 2 demonstrates the average cumulative regret vanishes when $f(\mathbf{x}^*) = h$, or reaches $\mid f(\mathbf{x}^*) - h \mid$ when the specified level does not exist for the function, with $\mathbf{x}^*$ being the set of points resulting in $f(\mathbf{x}^*)$ being the closest to h. In Theorem 2, cumulative regret is bounded as a function of the maximum information gain γ_t. The existing results in Theorem 5 of [15] provide an upper bound for γ_t for a squared exponential kernel with homogeneous length scale.

In Theorem 3 we provide a bound on the maximum information gain in the presence of a heterogeneous length scale as bounded by Theorem 1. Even with a heterogeneous length scale for the GP, it can be shown that maximum information gain remains bounded by the same order as that of a homogeneous length scale. This is described in Theorem 3 and provides guarantees for Theorem 2 under a non-stationary GP.

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* $\mathcal{O}((\log T)^{d+1})$.

Proof. Proof of Theorem 3 is provided in supplementary material.

Note in Theorem 3, regret bounds are of order $\mathcal{O}(T)$ only when the function does not have a level at h i.e. either $h > f_{max}$ or $h < f_{min}$. When $f_{min} \leq h \leq f_{max}$, then this term will go to 0, leaving an order of $\mathcal{O}(\sqrt{T(logT)^{d+1}})$.

2.4 Tuning Warping Hyper-parameters

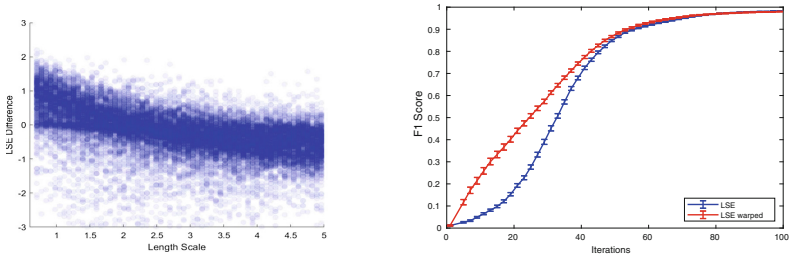
Before warping, the hyper-parameters $\boldsymbol{l}_0$ and $\boldsymbol{l}_1$ must be estimated. This can be done by separating a set of observations into training and test set. $\boldsymbol{l}_0$ and $\boldsymbol{l}_1$ are optimised for values which, when Algorithm 1 is applied using the training set, classification produces highest F1 score for the test set, after a pre-defined number of iterations.

3 Experimental Results

Comparison of the performance of the acquisition function with (our approach) and without (existing approach [8]) input warping was examined against one synthetic function and two real world problems. Our evaluation method uses a separate test set to measure classification accuracy into the super-level and sub-level sets. F1 scores are reported with error bars indicating standard error. All experiments were randomly initialised and run 50 times. Code for first experiment can be found at: https://bit.ly/37gNhPZ.

3.1 Mishra's Bird Function

In this experiment, we intend to find a super-level set of the Mishra's Bird benchmark function of the form $f(x_1, x_2) = sin(x_1)e^{(1-cos(x_2))^2} + cos(x_2)e^{(1-sin(x_1))^2} + (x_1 - x_2)^2$ at $h = 10$. At this level there are multiple disconnected regions. For the search, the length scale was warped with values for $\boldsymbol{l}_0$ and $\boldsymbol{l}_1$ being $[0.02, 0.38]$ and $[1.32, 0.002]$ respectively. The value of ϵ was 0.1 and $\sqrt{\beta_t} = \sqrt{\nu\tau_t}$ as defined in [3] where $\tau_t = 2log(t^{d/2+2}\pi^2/3\delta)$ with $\delta = 0.01$ and $\nu = 1$.



(a) (b)

Fig. 3. a) Change in acquisition function vs length scale. Positive value indicates increase in acquisition function value. Increased acquisition function values at shorter lengths vs longer length scales. b) F1 vs iteration results for classification of Bird Function at threshold $h = 10$

Figure 3a) illustrates experimentally the impact of warping the length scale on the acquisition function. Positive difference indicates that values in the warped scenario are higher than in the no-warping scenario. Results show for smaller length scales, the acquisition function is increased, while at larger length scales, the value is decreased. Changing the acquisition function via warping increases (or decreases) the chance of a point being selected as the next sample point.

Figure 3b) shows the comparative performance over 50 randomly initialised trials with and without warping of the search space. The classification accuracy and rate are improved notably with input warping accompanying the acquisition function.

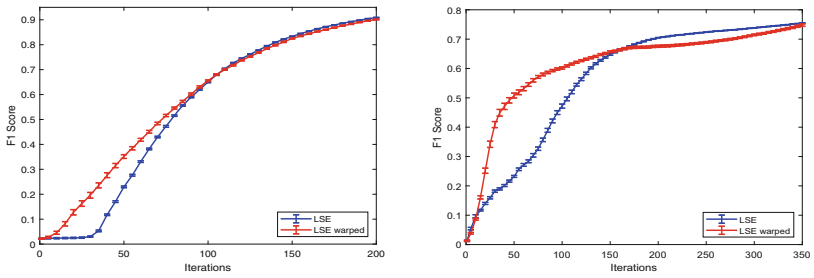
3.2 Car Crashworthiness Design

LS-DYNA, is a finite element modelling program which simulates complex scenarios in the physical world [9]. Using a simplified car crash simulation, we demonstrate an important application of level set estimation for the design of safe vehicles. The problem focuses on a vehicle moving at a constant velocity and crashing into a pole, resulting in the front of the car deforming. A car must be designed to maintain the safety of passengers. In both experiments below, the input parameters represent the mass of various car components. Altering these inputs alters the rigidity of the car. If too rigid, the passengers will experience injury from the forces of impact (eg. whiplash). If not rigid enough, the front of the car may crush and intrude into the passenger space. The objective of such a problem is to maximise the "crashworthiness" of the vehicle.

Experiment 1. In the first crash experiment there are two design parameters: mass of the front bumper bar *tbumper* and, mass of the front, hood and underside of the bonnet *thood*. Both range between 1 and 5, representing the thickness of the component. To construct the dataset, each of the two input parameters were sampled within the entire range in steps of 0.1, resulting in 1681 combinations. The output of the simulation is a Head injury criterion (*HIC*) with the objective being to maintain $HIC < 250$.

Experiment 2. In this scenario, the number of inputs is 6, representing thickness of hood *thood*, grill *tgrill*, roof *troof*, bumper *tbumper*, front of rails *trailf* and back of rails *trailb*. Inputs were sampled over a grid of 15,625 points. The output is frequency of car torsional vibration. The objective is to maintain torsional mode frequency < 1.9 Hz.

Figure 4 shows the comparative results.



(a) Experiment 1 results (b) Experiment 2 results

Fig. 4. (a) F1 score vs iteration for HIC < 250. LSE with warping outperforms standard LSE in initial iterations. LSE slightly outperforms LSE warped in final stage. This is considered negligible. (b) F1 score vs iteration for torsional mode frequency < 1.9 Hz. LSE with warping outperforming LSE considerably in early stages.

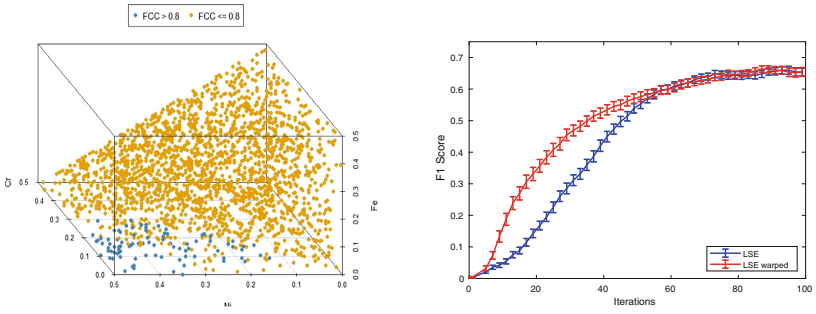
3.3 Ductile Alloy Design

Design of high entropy alloys (HEA) with exceptional physical properties is an active research area in the material science community. To assist in the design of such alloys, many practitioners use the High Entropy Alloy Database (TCHEA) on the Thermo-Calc software. Thermo-Calc is a powerful tool in Computational Thermodynamics and is popular for thermochemical calculations of heterogeneous phase equilibria and multicomponent phase analysis [10].

In this experiment we utilise the TCHEA database in Thermo-Calc for the design of 4 element alloy systems consisting of Iron (Fe), Nickle (Ni), Cobalt (Co) and Chromium (Cr). The objective is to determine the set of alloy compositions that, when cast at room temperature (27°), resulted in an Face-Centered Cubic (FCC) proportion of at least 80%. The input space was constrained such that the four element's mass percentage could range between 0–50%, and the sum of the elements must equal 100%. Figure 5a) shows the target region.

Due to dependent nature of input variables (by constraint that sum is 100%), only 3 elements were used. Figure 5b) shows the comparative results for 50 trials.

For most experiments classification from warping is faster in early iterations before converging, demonstrating the exploitative behaviour of the acquisition function from warping. This justifies the use of warping for level set estimation, particularly when function evaluations are expensive and budget limits the number of samples.



(a) (b)

Fig. 5. a) FCC vs elemental compositions. Regions where FCC $\geq$ 80% indicated in blue. b) Results for classification of alloy with threshold at 80% FCC. LSE with warping outperforms standard LSE before converging to same rate in later iterations. (Color figure online)

3.4 Computational Time

Computational impact of warping comes from constructing the complex covariance matrix in the non-stationary GP. The computational efficiency of being able to vectorise the covariance matrix construction with a constant length scale is not possible in the changing length scale scenario and for loops are needed. For example, run time for 50 iterations of Mishra's Bird function without warping is on average 12 s, whilst with the warping, time is around 5 min. However, it is assumed that function evaluation time for the real world cases are well above the optimiser's run time.

4 Conclusion

This paper presented a novel means in which a complex covariance function can be constructed by distorting the length scale of the GP from which the acquisition function samples from. By doing so, areas with a high potential for being at the level are expanded, thereby increasing the chance of sampling in these regions. Conversely, areas with lower potential are contracted. The warping metric valued the same characteristics as the acquisition function, allowing the two to operate together. The warping metric however results in the acquisition function behaving more exploitatively, which is beneficial in level set estimation problems. Guarantees of convergence were presented as well as bounds on the length scale range and maximum information gain.

Acknowledgement. The authors would like to acknowledge and thank Dr Huong Ha and Dr Stewart Greenhill for their contributions to the proofs and experimental section in the paper, respectively. This research was partially funded by the Australian Government through the Australian Research Council (ARC). Prof Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

References

1. Bect, J., Ginsbourger, D., Li, L., Picheny, V., Vázquez, E.: Sequential design of computer experiments for the estimation of a probability of failure. Stat. Comput. **22**(3), 773–793 (2012). https://doi.org/10.1007/s11222-011-9241-4
2. Bogunovic, I., Scarlett, J., Krause, A., Cevher, V.: Truncated variance reduction: a unified approach to Bayesian optimization and level-set estimation. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016, pp. 1515–1523. Curran Associates Inc., USA (2016)
3. Brochu, E., Cora, V.M., de Freitas, N.: A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. CoRR abs/1012.2599 (2010). http://arxiv.org/abs/1012.2599
4. Bryan, B., Nichol, R.C., Genovese, C.R., Schneider, J., Miller, C.J., Wasserman, L.: Active learning for identifying function threshold boundaries. In: Weiss, Y., Schölkopf, B., Platt, J.C. (eds.) Advances in Neural Information Processing Systems, vol. 18, pp. 163–170. MIT Press (2006)
5. Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., Richet, Y.: Fast parallel kriging based stepwise uncertainty reduction with application to the identification of an excursion set. Technometrics **56**(4), 455–465 (2014)
6. Eppinger, R., Kuppa, S., Saul, R., Sun, E.: Supplement: development of improved injury criteria for the assessment of advanced automotive restraint systems: Ii (2000)
7. Garg, A., et al.: Tumor localization using automated palpation with Gaussian process adaptive sampling. In: 2016 IEEE International Conference on Automation Science and Engineering (CASE), pp. 194–200. IEEE (2016)
8. 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, IJCAI 2013, pp. 1344–1350. AAAI Press (2013)
9. Hallquist, J.O., Manual, L.D.T.: Livermore software technology corporation. Livermore, CA (1998)

10. Andersson, J.-O., Helander, T., Höglund, L., Shi, P., Sundman, B.: Thermo-Calc and DICTRA, computational tools for materials science. Calphad **26**, 273–312 (2002)
11. Paciorek, C.J., Schervish, M.J.: Nonstationary covariance functions for Gaussian process regression. In: Thrun, S., Saul, L.K., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems, vol. 16, pp. 273–280. MIT Press (2004)
12. Plagemann, C., Kersting, K., Burgard, W.: Nonstationary Gaussian process regression using point estimates of local smoothness. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS (LNAI), vol. 5212, pp. 204–219. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87481-2_14
13. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning. MIT Press, Cambridge (2006)
14. 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, ICML 2014, vol. 32, pp. II-1674–II-1682 (2014). JMLR.org
15. Srinivas, N., Krause, A., Kakade, S., Seeger, M.: Gaussian process optimization in the bandit setting: no regret and experimental design. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML 2010, pp. 1015–1022. Omnipress, USA (2010)
16. Zanette, A., Zhang, J., Kochenderfer, M.J.: Robust super-level set estimation using Gaussian processes. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11052, pp. 276–291. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10928-8_17