Robust Transfer Learning for Active Level Set Estimation with Locally Adaptive Gaussian Process Prior

Giang Ngo* G.NGO@DEAKIN.EDU.AU
Dang Nguyen D.NGUYEN@DEAKIN.EDU.AU
Sunil Gupta SUNIL.GUPTA@DEAKIN.EDU.AU
Applied Artificial Intelligence Institute (A2I2), Geelong, Victoria, Australia

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

The objective of active level set estimation for a black-box function is to precisely identify regions where the function values exceed or fall below a specified threshold by iteratively performing function evaluations to gather more information about the function. This becomes particularly important when function evaluations are costly, drastically limiting our ability to acquire large datasets. A promising way to sample-efficiently model the black-box function is by incorporating prior knowledge from a related function. However, this approach risks slowing down the estimation task if the prior knowledge is irrelevant or misleading. In this paper, we present a novel transfer learning method for active level set estimation that safely integrates a given prior knowledge while constantly adjusting it to guarantee a robust performance of a level set estimation algorithm even when the prior knowledge is irrelevant. We theoretically analyze this algorithm to show that it has a better level set convergence compared to standard transfer learning approaches that do not make any adjustment to the prior. Additionally, extensive experiments across multiple datasets confirm the effectiveness of our method when applied to various different level set estimation algorithms as well as different transfer learning scenarios.

Keywords: Level set estimation; Active learning; Transfer learning

1. Introduction

The level set estimation (LSE) problem involves classifying whether $\mathbf{x}$ belongs to the superlevel set (i.e. $f(\mathbf{x})$ exceeds a threshold h) or the sublevel set (i.e. $f(\mathbf{x})$ is below h). LSE arises in many important applications including environmental monitoring (Gotovos et al. (2013)), computational chemistry (Tran et al. (2021)), and material characterization (Hozumi et al. (2023)). Two common difficulties of these applications are that the underlying function relating its input variables to the output is unknown, i.e. a *black-box* function, and that evaluating such black-box functions requires conducting real-world experiments that are often very expensive. To model the black-box function for level set classification in such scenarios, an intuitive approach is to use an active learning method to sequentially acquire new data points that are highly informative for determining the level set. The focus of active LSE algorithms is often on finding the locations of those informative data points at each iteration. This is done by optimizing an *acquisition function*, which is an indicator of how useful a data point is in classifying the level sets. The acquisition function is constructed using a surrogate model of the black-box function built with the existing function

* corresponding author

© 2024 G. Ngo, D. Nguyen & S. Gupta.

evaluations. The black-box function is then evaluated at the next point suggested by the acquisition function, and this new observation is used to update the surrogate model. The surrogate model will be used to classify the level set of any point either at every iteration or when an evaluation budget runs out.

Due to the high cost of function evaluations, leveraging additional relevant information can significantly reduce the number of evaluations required to achieve the desired estimation accuracy. Such relevant information may come from prior information based on domain knowledge or data from a related LSE task completed in the past. For instance, in the context of *algorithmic assurance* (Gopakumar et al. (2018)), where the objective is to identify scenarios in which the performance of a pretrained machine learning model reaches a certain threshold, prior information can be sourced from the identified scenarios of a related model. Other examples include the elicitation of expert knowledge to form a prior function to model the output of complex computer code by Oakley (2002) or the problem of discovering Higgs boson when Golchi (2016) suggested the construction of a prior mean function by either Monte Carlo studies or a polynomial fit to simulated data.

The previous works on using prior information for active LSE are limited. Only Hozumi et al. (2023) attempted the problem but assumed that data from a related task were available. There have been plenty of attempts to use prior information in a related problem known as Bayesian Optimization (BO), where the prior information has been used to accelerate the finding of the optimum of a function (Swersky et al. (2013); Shilton et al. (2017); Wistuba et al. (2015)). However, since the function at hand is black-box, assessing the relevance of prior information is not always straightforward. At times, it can be misleading and hinder the LSE task. Though the prior information is believed to be related to f, mismatches can occur, either entirely or partially across the input space. For example, in environmental monitoring, the condition of the whole environment or part of it can change compared to the previous LSE task. Existing LSE algorithms often employ Gaussian Process (GP) (Rasmussen and Williams (2005)) as the surrogate model in which the posterior mean prediction at $\mathbf{x}$ can combine information from existing data points and the prior distribution. In case of mismatches, the posterior mean that uses prior information will be less accurate than the one without it.

This inaccuracy in modeling f will weaken the sample efficiency and accuracy of an LSE algorithm. Firstly, it will be harder for the acquisition function to correctly indicate how informative a point is in level set classification. One popular criterion for acquisition functions is to prioritize data points that are near the threshold boundary (Bryan et al. (2005); Zanette et al. (2019); Gotovos et al. (2013)), and this becomes non-trivial if the acquisition function fails to tell if the function value at a point is close to h or not due to inaccurate function estimation. The algorithm then needs to explore more to locate the targeted points. Another obvious consequence is inaccurate level set classification where the function is not estimated accurately, lowering the accuracy of the LSE algorithm. The GP posterior will need more data points to mitigate the impact of inaccurate prior information in a region, which again hurts the sample efficiency of the algorithm.

Given the above issues, it is necessary to control the use of prior information so that it will not adversely affect the LSE algorithm in regions where it is observed to be irrelevant to f. However, we would still like to use the prior information in regions with insufficient data points. This paper proposes a *locally adaptive prior* that adjusts the prior function

according to the mismatches between itself and the observed function evaluations. The adjusted prior becomes similar to the observed function values in the discovered regions and reverts to its own form in undiscovered regions, allowing a robust transfer of prior knowledge into an LSE algorithm regardless of any potential unseen mismatch between the prior knowledge and the actual underlying function. Thus, the prior information is safely transferred into the surrogate model without letting the observed mismatches affect the main LSE algorithm.

Our main contributions are:

1. A novel method to safely transfer prior information in active LSE to accelerate the estimation of level sets for a black-box function.
2. A theoretical analysis of the convergence of our method as well as its efficacy compared to transfer learning using priors without any adaptations.
3. Extensive empirical evidence showing robust performance of our method on multiple LSE tasks, on different LSE algorithms, with different levels of mismatches.

2. Related Works

Level Set Estimation

The main difference among LSE algorithms is in the selection of a new data point to evaluate its function value at each iteration. The selected point needs to reveal as much information about the level sets of f as possible. A popular direction in active LSE is to select the point where the surrogate model is the most uncertain in performing level set classification. Bryan et al. (2005) proposed maximizing the Straddle heuristic $1.96\sigma_t(\mathbf{x}) - |\mu_t(\mathbf{x}) - h|$ where $\mu_t(\mathbf{x})$ and $\sigma_t(\mathbf{x})$ are the posterior mean and standard deviation of a GP surrogate model for f at iteration t. This acquisition function prioritizes a point either in the undiscovered region or near the threshold boundary. ActiveLSE (Gotovos et al. (2013)) generalized the Straddle heuristic by maximizing the classification ambiguity $\min\{\max(C_t(\mathbf{x}) - h, h - \min(C_t(\mathbf{x})))\}$ in which $C_t(\mathbf{x})$ is the successive intersection of the confidence intervals of $f(\mathbf{x})$ at each iteration. Due to the maintenance of $C_t(\mathbf{x})$, ActiveLSE needs to operate on a fixed discrete number of data points while the Straddle heuristic can work on a continuous domain, and the two algorithms often share similar empirical performance. C2LSE (Ngo et al. (2024)) attempted to select points with the smallest classification confidence by maximizing $\sigma_t(\mathbf{x})/\max(\epsilon, |\mu_t(\mathbf{x}) - h|)$ where $\epsilon > 0$ prevents the algorithm from getting stuck at the boundary and also controls the algorithm's accuracy.

Another direction of active LSE is related to one-step lookahead mechanism. TruVar (Bogunovic et al. (2016)) chose a point such that it minimizes the sum of truncated variances over a set of fixed points that arise from choosing that point. This algorithm also considers point-wise evaluation costs to favor locations with cheaper evaluations. RMILE (Zanette et al. (2019)) is another algorithm within this category that is influenced by the expected improvement algorithm. Its acquisition function calculates the expected change in volume of the superlevel set if a point $\mathbf{x}$ is chosen and also takes into account the uncertainty at a point to avoid getting stuck. Other approaches for active LSE include hierarchical space partitioning (Shekhar and Javidi (2019)), information-theoretic acquisition (Nguyen et al. (2021)), and using experimental design (Mason et al. (2022)).

Transfer Learning for Level Set Estimation
There are limited works in transfer learning for active LSE. Hozumi et al. (2023) used an active LSE approach to identify defective areas on silicon ingots for solar cell production where they utilized measurements on a previous ingot by employing Diff-GP (Shilton et al. (2017)), which is originally designed for BO to incorporate data from a source task into the target task by modeling the difference between the two tasks. BO is often seen as related to active LSE where its goal is to find the optimum of a black-box function, and BO also often employs a GP as its surrogate model. Existing works in transfer learning for BO include jointly learned kernel in multi-task setting (Swersky et al. (2013); Poloczek et al. (2016)), seeing source data as noisy observation of the target task (Shilton et al. (2017); Theckel Joy et al. (2019)), and search space design (Wistuba et al. (2015); Perrone et al. (2019)). These works also consider the setting where prior information is from available source data, so they are not applicable to the setting where prior information is from domain knowledge.

3. Proposed Method

We propose a novel method to integrate prior information into the active LSE process. Our main idea is to adapt the prior locally based on any observed discrepancies before using it for the active LSE.

3.1. Problem Formulation

We now formally define the active LSE problem when *prior knowledge is available*.
Active level set estimation problem: The goal is to iteratively obtain new data points of a black-box function $f : \mathcal{X} \to \mathbb{R}$ with compact set $\mathcal{X} \in \mathbb{R}^d$ to improve a classifier which accurately determines whether any point $\mathbf{x} \in \mathcal{X}$ is in the superlevel set $H = \{\mathbf{x} \in \mathcal{X} | f(\mathbf{x}) > h\}$ or the sublevel set $L = \{\mathbf{x} \in \mathcal{X} | f(\mathbf{x}) < h\}$ given a threshold $h \in \mathbb{R}$.
Using prior knowledge: We assume that the prior knowledge about f is in the form of $u_p : \mathcal{X} \to \mathbb{R}$. This prior knowledge may be available through a domain-specific simulator, approximate models, or as a result of fitting a machine learning model on the data from a related setting. Our goal is to leverage u_p in active LSE for f such that we either need fewer data points or achieve more accurate classification than starting active LSE from scratch.

3.2. Prior in Gaussian Process

GPs are often used as the surrogate models for f in the problem of active LSE. A GP is specified by its mean function $\mu(\mathbf{x})$ and its kernel $k(\mathbf{x}, \mathbf{x}')$. The posterior distribution over f of a GP given its zero mean prior and noisy observations is as follows:

$$\mu_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^T (\mathbf{K}_t + \sigma^2 \mathcal{I})^{-1} \mathbf{Y}_t \tag{1}$$

$$\sigma_t^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^T (\mathbf{K}_t + \sigma^2 \mathcal{I})^{-1} \mathbf{k}_t(\mathbf{x}) \tag{2}$$

where $\mathbf{k}_t(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_i)]_{i=1}^t$, $\mathbf{K}_t = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^t$, and $\mathbf{Y}_t = f(\mathbf{X}_t) + \mathbf{E}_t$ are the noisy observations of f at $\mathbf{X}_t = [\mathbf{x}_i]_{i=1}^t$ with observation noises $\mathbf{E}_t = [\eta_i]_{i=1}^t$, $\eta_i \sim \mathcal{N}(0, \sigma^2)$.

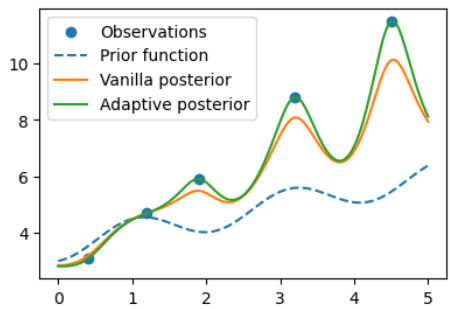

Figure 1: Blue: prior function; Orange: posterior mean by vanilla transfer learning; Green: posterior mean with our method AP-LSE. For the two data points on the left when the prior is still relevant compared to the observed function values, the posterior mean obtained with vanilla transfer learning still closely matches the observed values. When the difference is large as seen on the right, $\bar{\mu}_t$ starts to deviate from the data points. On the other hand, the posterior mean obtained with AP-LSE remains better fitted in both situations.

A non-trivial prior function u_p can be incorporated in the posterior mean distribution of such Gaussian Processes by seeing it as the prior mean function as follows:

$$\bar{\mu}_t(\mathbf{x}) = u_p(\mathbf{x}) + \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathcal{I})^{-1}(\mathbf{Y}_t - u_p(\mathbf{X}_t)) \tag{3}$$

Directly using u_p as in Eq.3, which we refer to as *vanilla transfer learning* for GP, can help to transfer possibly related knowledge into the posterior mean of the Gaussian Process. However, it is possible that with the observations of f, one may discover differences between f and u_p at certain points in $\mathbf{X}_t$, leading to inaccurate estimation of f due to the bias introduced by u_p as seen in Figure 1. As explained in Section 1, mitigating the effect of u_p in the posterior mean around these points is necessary for active LSE tasks.

3.3. Locally adaptive prior for active level set estimation

The challenge is in adjusting u_p around observed mismatching locations but still transferring prior knowledge into the GP at other locations where we have no data. A simple solution is to define a region in $\mathcal{X}$ that surrounds the observed data points where the prior function u_p is not used, and the posterior mean becomes μ_t as in Eq.1 instead. However, such a region will disrupt the smoothness of the posterior mean and likely make acquisition optimization difficult. It is also challenging to properly define the region since there is little information to determine the effective radius (i.e. how far away from $\mathbf{X}_t$ is sufficient).

To address this challenge, we propose **Adaptive Prior Level Set Estimation** (AP-LSE) which smoothly adjusts the prior function u_p by an additive term u_{dt} according to the observed difference between the current observations and the prior function, i.e. $\mathbf{Y}_t - u_p(\mathbf{X}_t)$, before the posterior mean is used for computing acquisition value and level set classification. The adjustment u_{dt} is an interpolation of the difference $\mathbf{Y}_t - u_p(\mathbf{X}_t)$. Specifically, the new posterior mean with adjusted prior is given as:

$$\tilde{\mu}_t(\mathbf{x}) = u_p(\mathbf{x}) + u_{dt}(\mathbf{x}) + \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathcal{I})^{-1}(\mathbf{Y}_t - u_p(\mathbf{X}_t) - u_{dt}(\mathbf{X}_t)) \tag{4}$$

where $u_{dt} = \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathcal{I})^{-1}(\mathbf{Y}_t - u_p(\mathbf{X}_t))$. The value of $u_{dt}(\mathbf{x})$ reflects the adjustment of the prior function according to $\mathbf{Y}_t$. When $\mathbf{x}$ is close to points in $\mathbf{X}_t$ (i.e. this region has

been sampled), the posterior mean will favor the actual data observations over the prior knowledge. When $\mathbf{x}$ is not close to points in $\mathbf{X}_t$ (i.e. there is still a lot of uncertainty in this region), $u_{dt}(\mathbf{x})$ will be closer to 0 which means $u_p(\mathbf{x}) + u_{dt}(\mathbf{x})$ becomes similar to $u_p(\mathbf{x})$.

Algorithm 1: Adaptive Prior Level Set Estimation

Input: prior function u_p, threshold h, budget T.
$D_0 \leftarrow \emptyset; t \leftarrow 1;$
while $t < T$ **do**
1. Fit the hyperparameters of k_t with $\mathcal{GP}(u_p, k(\mathbf{x}, \mathbf{x}'))$ and observations $[\mathbf{x}_i, y_i]_{i=1}^t$;
2. Obtain the adjusted prior function: $u_{p,t}(\mathbf{x}) = u_p(\mathbf{x}) + u_{dt}(\mathbf{x})$;
3. Obtain the posterior mean function: $\tilde{\mu}_t(\mathbf{x}) = u_{p,t}(\mathbf{x}) + \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathcal{I})^{-1}(\mathbf{Y}_t - u_{p,t}(\mathbf{X}_t))$;
4. Select next evaluation location $\mathbf{x}_t$ by optimizing a level set acquisition function a_t over $\mathcal{X}$ (see examples of a_t below);
5. Evaluate f at $\mathbf{x}_t$: $y_t = f(\mathbf{x}_t) + \eta_i$;
6. $D_t = D_{t-1} \cup (\mathbf{x}_t, y_t)$;
7. $t = t + 1$

end
Output: predicted level set classification;

Details of AP-LSE are in Algorithm 1. Each iteration first fits kernel hyperparameters using the current observations of f and uses the fitted kernel to obtain the adjustment function u_{dt}. The posterior mean is computed with the fitted kernel and the adjusted prior function $u_{p,t}(\mathbf{x}) = u_p(\mathbf{x}) + u_{dt}(\mathbf{x})$. The location of the next observation is obtained by optimizing an acquisition function $a_t(\mathbf{x})$. Examples of such acquisition functions include:

- The Straddle heuristic (Bryan et al. (2005)): $a_t(\mathbf{x}) = 1.96\sigma_t(\mathbf{x}) - |\tilde{\mu}_t(\mathbf{x}) - h|$
- C2LSE (Ngo et al. (2024)): $a_t(\mathbf{x}) = \frac{\sigma_t(\mathbf{x})}{\max(\epsilon, |\tilde{\mu}_t(\mathbf{x}) - h|)}$ with $\epsilon > 0$
- RMILE (Zanette et al. (2019)): $a_t(\mathbf{x}) = \max(\mathbb{E}_{y^+}|I^+| - |I^\epsilon|, \lambda\sigma_t(\mathbf{x}))$ for $\epsilon > 0, \lambda > 0$, with $|I^\epsilon| = \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{1}(P(f(\mathbf{x}) > h - \epsilon) > \delta)$ and $|I^+| = \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{1}(P^+(f(\mathbf{x}) > h) > \delta)$ where $P+$ is calculated based on the GP posterior if we evaluate $f(\mathbf{x})$ and observe value y^+.

The algorithm then evaluates f at the selected location $\mathbf{x}_t$ to obtain the noisy observation y_t, and the new data point $(\mathbf{x}_t, y_t)$ joins the existing ones to form the training data D_t. The predicted level set classification can be obtained either at the end of each iteration or when the algorithm terminates by applying classification rules. An example of such rules are:

- Predicted superlevel set $\hat{H}_t = \{\mathbf{x} \in \mathcal{X} : \tilde{\mu}_t(\mathbf{x}) - \beta\sigma_t(\mathbf{x}) > h\}$
- Predicted sublevel set $\hat{L}_t = \{\mathbf{x} \in \mathcal{X} : \tilde{\mu}_t(\mathbf{x}) + \beta\sigma_t(\mathbf{x}) < h\}$
- Unclassified set $\hat{M}_t = \{\mathbf{x} \in \mathcal{X} : \tilde{\mu}_t(\mathbf{x}) - \beta\sigma_t(\mathbf{x}) \leq h \leq \tilde{\mu}_t(\mathbf{x}) + \beta\sigma_t(\mathbf{x})\}$

with confidence parameter β.

AP-LSE provides a transfer learning method suitable for most LSE algorithms where $\tilde{\mu}_t$ helps us mitigate the use of u_p when we have sufficient knowledge about f while still allowing the leverage of the prior information when there is no data about f.

4. Theoretical Analysis

Our goal is to show the effectiveness of using $\tilde{\mu}_t$ over $\bar{\mu}_t$ for active LSE. The analysis first shows that by using $\tilde{\mu}_t$, the GP posterior better fits the training data points. We show the upper bounds for the generalization errors of $\tilde{\mu}_t$ and $\bar{\mu}_t$ and finally a better level set classification confidence using the former.

4.1. Fitting Error

One design goal of AP-LSE is for the posterior mean to favor the observations at points close to $\mathbf{X}_t$. This section shows that $\tilde{\mu}_t$ is closer to the observed values $\mathbf{Y}_t$ at the observed locations compared to $\bar{\mu}_t$, which is demonstrated by a better fitting mean squared error.

Proposition 1 *Let $\alpha_t(\boldsymbol{x}) = \boldsymbol{k}_t(\boldsymbol{x})^T(\boldsymbol{K}_t + \sigma^2\mathcal{I})^{-1}$, $Z_N = [z_i]_{i=1}^t$ be a vector of constants, and $\mathcal{H}_k(\mathcal{X})$ be a RKHS corresponding to $k(\boldsymbol{x}, \boldsymbol{x}')$ on $\mathcal{X}$. The following inequality holds:*

$$\sum_{i=1}^{t}(z_i - \alpha_t(\boldsymbol{x}_i)Z_t)^2 \leq \sum_{i=1}^{t} z_i^2 \tag{5}$$

Proof *Theorem 3.4 by Kanagawa et al. (2018) stated that $f^*(\boldsymbol{x}) = \alpha_t(\boldsymbol{x})Z_t$ is the solution of the following regularized least-squares problem:*

$$\hat{f}(\boldsymbol{x}) = \underset{f \in \mathcal{H}_k}{argmin} \frac{1}{t}\sum_{i=1}^{t}(f(\boldsymbol{x}_i) - y_i)^2 + \frac{\sigma^2}{t}||f||_{\mathcal{H}_k}^2$$

Therefore, it follows that:

$$\frac{1}{t}\sum_{i=1}^{t}(f^*(\boldsymbol{x}_i) - y_i)^2 + \frac{\sigma^2}{t}||f^*||_{\mathcal{H}_k}^2 \leq \frac{1}{t}\sum_{i=1}^{t}(g(\boldsymbol{x}_i) - y_i)^2 + \frac{\sigma^2}{t}||g||_{\mathcal{H}_k}^2 \text{ for } g(\boldsymbol{x}) = 0. \tag{6}$$

This means $\sum_{i=1}^{t}(z_i - \alpha_t(\boldsymbol{x}_i)Z_t)^2 + \sigma^2||\alpha_t(\boldsymbol{x})Z_t||_{\mathcal{H}_k}^2 \leq \sum_{i=1}^{t} z_i^2$, or the inequality holds. ■

The following result states that AP-LSE fits the training data points better than vanilla transfer learning.

Theorem 2 *Let $\mathcal{D}_t = [\boldsymbol{x}_i, y_i]_{i=1}^t$ be the observations of f. The fitting error of $\tilde{\mu}_t(\boldsymbol{x})$ is upper bounded by that of $\bar{\mu}_t(\boldsymbol{x})$, which is bounded by the difference between u_p and f. That is:*

$$\sum_{i=1}^{t}(\tilde{\mu}_t(\boldsymbol{x}_i) - y_i)^2 \leq \sum_{i=1}^{t}(\bar{\mu}_t(\boldsymbol{x}_i) - y_i)^2 \leq \sum_{i=1}^{t}(u_p(\boldsymbol{x}_i) - y_i)^2 \tag{7}$$

Proof *We have that:*

$$\begin{aligned}
\tilde{\mu}_t(\boldsymbol{x}_i) - y_i &= u_p(\boldsymbol{x}_i) + u_{dt}(\boldsymbol{x}_i) + \alpha_t(\boldsymbol{x}_i)(\boldsymbol{Y}_t - u_p(\boldsymbol{X}_t) - u_{dt}(\boldsymbol{X}_t)) - y_i \\
&= u_p(\boldsymbol{x}_i) + \alpha_t(\boldsymbol{x}_i)(\boldsymbol{Y}_t - u_p(\boldsymbol{X}_t)) - y_i + \alpha_t(\boldsymbol{x}_i)(\boldsymbol{Y}_t - u_p(\boldsymbol{X}_t) - \alpha_t(\boldsymbol{X}_t)(\boldsymbol{Y}_t - u_p(\boldsymbol{X}_t))) \\
&= \bar{\mu}_t(\boldsymbol{x}_i) - y_i + \alpha_t(\boldsymbol{x}_i)(\boldsymbol{Y}_t - \bar{\mu}_t(\boldsymbol{X}_t)) \\
&= u(\boldsymbol{x}_i) - \alpha_t(\boldsymbol{x}_i)u(\boldsymbol{X}_t),
\end{aligned}$$

for $1 \leq i \leq t$ *with* $u(\boldsymbol{x}) = \bar{\mu}_t(\boldsymbol{x}) - y$, y *being the observation of* f *at* $\boldsymbol{x}$. *Using Proposition 1, we can establish the left inequality in (7) as follows:*

$$\sum_{i=1}^{t}(u(\boldsymbol{x}_i) - \alpha_t(\boldsymbol{x}_i)u(\boldsymbol{X}_t))^2 \leq \sum_{i=1}^{t} u(\boldsymbol{x}_i)^2 = \sum_{i=1}^{t}(\bar{\mu}_t(\boldsymbol{x}_i) - y_i)^2$$

Let $g(\boldsymbol{x}) = u_p(\boldsymbol{x}) - y$. *We then have*

$$u(\boldsymbol{x}_i) = u_p(\boldsymbol{x}_i) - y + \alpha_t(\boldsymbol{x}_i)(\boldsymbol{Y}_t - u_p(\boldsymbol{X}_t)) = g(\boldsymbol{x}) - \alpha_t(\boldsymbol{x})g(\boldsymbol{X}_t) \tag{8}$$

Using Proposition 1, we can establish the right inequality in (7):

$$\sum_{i=1}^{t} u(\boldsymbol{x}_i)^2 = \sum_{i=1}^{t}(g(\boldsymbol{x}_i) - \alpha_t(\boldsymbol{x})g(\boldsymbol{X}_t))^2 \leq \sum_{i=1}^{t} g(\boldsymbol{x}_i)^2$$

Both inequalities in (7) hold as a result. ■

4.2. Generalization Error

In addition to the fitting error at observed locations, it is more important that a model generalizes well at other locations. Here we give bounds to the generalization errors of $\tilde{\mu}_t$ and $\bar{\mu}_t$. Let us denote $\bar{e}_t(\mathbf{x}) = \bar{\mu}_{t-1}(\mathbf{x}) - f(\mathbf{x})$ and $\tilde{e}_t(\mathbf{x}) = \tilde{\mu}_{t-1}(\mathbf{x}) - f(\mathbf{x})$.

Lemma 3 *Let* $\rho_t = \sup_{\boldsymbol{x}\in\mathcal{X}} \inf_{\boldsymbol{x}_i \in \boldsymbol{X}_t} \|\boldsymbol{x} - \boldsymbol{x}_i\|_2$. *Let* y *be the observation of* f *at* $\boldsymbol{x}$. *Assume that:*

1. $f \in \mathcal{H}_k(\mathcal{X})$,
2. $\mathcal{H}_k(\mathcal{X})$ *is isomorphic to the Sobolev space* $\mathcal{H}^\tau(\mathcal{X})$ *for some* $\tau = n + r$, *with* $n \in \mathbb{N}, n > d/2$ *and* $0 < r < 1$,
3. $u_p, f \in \mathcal{H}^\tau(\mathcal{X})$,
4. *The Gaussian Process uses radial basis function (RBF) kernel.*

Then there exist a constant C *independent of* f, u_p, *and* t, *we have for* $0 \leq \xi \leq \tau$:

$$\|\bar{e}_t\|_{\mathcal{H}^\xi(\mathcal{X})} \leq C\rho_t^{\tau-\xi}\|u_p - f\|_{\mathcal{H}^\tau(\mathcal{X})} + \|\alpha_t \boldsymbol{E}_t\|_{\mathcal{H}^\xi(\mathcal{X})}, \tag{9}$$

$$\text{and } \|\tilde{e}_t\|_{\mathcal{H}^\xi(\mathcal{X})} \leq C\rho_t^{\tau-\xi}\|\bar{e}_t\|_{\mathcal{H}^\tau(\mathcal{X})} + \|\alpha_t \boldsymbol{E}_t\|_{\mathcal{H}^\xi(\mathcal{X})}, \tag{10}$$

Proof *We have that:*

$$\begin{aligned}\|\bar{\mu}_{t-1} - f\|_{\mathcal{H}^\xi(\mathcal{X})} &= \|u_p - f - \alpha_t(u_p(\boldsymbol{X}_t) - \boldsymbol{Y}_t)\|_{\mathcal{H}^\xi(\mathcal{X})} \\ &\leq \|u_p - f - \alpha_t(u_p(\boldsymbol{X}_t) - f(\boldsymbol{X}_t))\|_{\mathcal{H}^\xi(\mathcal{X})} + \|\alpha_t \boldsymbol{E}_t\|_{\mathcal{H}^\xi(\mathcal{X})}\end{aligned} \tag{11}$$

Given the assumptions, we have the following using Lemma 4.1 in Narcowich et al. (2006):

$$\|u_p - f - \alpha_t(u_p(\boldsymbol{X}_t) - f(\boldsymbol{X}_t))\|_{\mathcal{H}^\xi(\mathcal{X})} \leq C\rho_t^{\tau-\xi}\|u_p - f\|_{\mathcal{H}^\tau(\mathcal{X})} \tag{12}$$

Combining Ineq.12 with Ineq.11 establishes Ineq.9. Ineq.10 can be established similarly. ■

The first assumption is a general one on f being from the native space $\mathcal{H}_k(\mathcal{X})$, similar to Srinivas et al. (2010). The second and third assumptions mean that both the prior function u_p and the true underlying function f have similar smoothness τ in the Sobolev space $\mathcal{H}^\tau(\mathcal{X})$ which is equal to the native space $\mathcal{H}_k(\mathcal{X})$. The last assumption limits the following theoretical results to only LSE algorithms that employ an RBF kernel for the GP, but our experiments show AP-LSE is still effective with other kernels.

Lemma 4 *When applied to an LSE algorithm that eventually evaluates every point in an arbitrarily dense finite grid, the generalization errors of both vanilla transfer learning and AP-LSE converge to zero, that is*

$$\lim_{t\to\infty} |\bar{e}_t(\boldsymbol{x})| = 0 \text{ and } \lim_{t\to\infty} |\tilde{e}_t(\boldsymbol{x})| = 0. \tag{13}$$

In addition, let $\bar{r}$ and $\tilde{r}$ be the convergence rates for the upper bounds of $\{\|\bar{e}_t\|_{\mathcal{H}^\xi(\mathcal{X})}\}_{t=1}^\infty$ and $\{\|\tilde{e}_t\|_{\mathcal{H}^\xi(\mathcal{X})}\}_{t=1}^\infty$ in Lemma 3 respectively. Then $\bar{r} < \tilde{r}$.

Proof *Examples of LSE algorithms that eventually evaluates every point in an arbitrarily dense finite grid include C2LSE and RMILE (as shown in the proof of Theorem 6 by Ngo et al. (2024) and Lemma 4 by Zanette et al. (2019), respectively). This means that $\boldsymbol{X}_t \to \mathcal{X}$ as $t \to \infty$ for a continuous domain $\mathcal{X}$, or equivalently $\lim_{t\to\infty} \rho_t = 0$. The first consequence is that $\lim_{t\to\infty} C\rho_t^{\tau-\xi}\|u_p - f\|_{\mathcal{H}^\tau(\mathcal{X}_{\rho,t})} = 0$.*

Secondly, Chowdhury and Gopalan (2017) showed that:

$$|\alpha_t(\boldsymbol{x})\boldsymbol{E}_t| \le \sigma^{-1/2}\sigma_{t-1}(\boldsymbol{x})\sqrt{\boldsymbol{E}_t^T K_t(K_t + \sigma^2\mathcal{I})^{-1}\boldsymbol{E}_t}$$

If $\boldsymbol{X}_t \to \mathcal{X}$ as $t \to \infty$, we have $\lim_{t\to\infty}|\sigma_t(\boldsymbol{x})| = 0$, so $\lim_{t\to\infty}|\alpha_t(\boldsymbol{x})\boldsymbol{E}_t| = 0$ and eventually, $\lim_{t\to\infty}\|\alpha_t\boldsymbol{E}_t\|_{\mathcal{H}^\xi(\mathcal{X})} = 0$. As a result, we have that $\lim_{t\to\infty}\|\bar{\mu}_t - f\|_{\mathcal{H}^\xi(\mathcal{X}_{\rho,t})} = 0$, and consequently, $\lim_{t\to\infty}|\bar{e}_t(\boldsymbol{x})| = 0$. The convergence of AP-LSE's error can be established similarly.

Regarding convergence rates, we have that:

$$\bar{r} = \lim_{t\to\infty} \frac{C\rho_{t+1}^{\tau-\xi}\|u_p - f\|_{\mathcal{H}^\tau(\mathcal{X})} + \|\alpha_{t+1}\boldsymbol{E}_{t+1}\|_{\mathcal{H}^\xi(\mathcal{X})}}{C\rho_t^{\tau-\xi}\|u_p - f\|_{\mathcal{H}^\tau(\mathcal{X})} + \|\alpha_t\boldsymbol{E}_t\|_{\mathcal{H}^\xi(\mathcal{X})}}, \tag{14}$$

$$\text{and } \tilde{r} = \lim_{t\to\infty} \frac{C\rho_{t+1}^{\tau-\xi}\|\bar{e}_{t+1}\|_{\mathcal{H}^\tau(\mathcal{X})} + \|\alpha_{t+1}\boldsymbol{E}_{t+1}\|_{\mathcal{H}^\xi(\mathcal{X})}}{C\rho_t^{\tau-\xi}\|\bar{e}_t\|_{\mathcal{H}^\tau(\mathcal{X})} + \|\alpha_t\boldsymbol{E}_t\|_{\mathcal{H}^\xi(\mathcal{X})}} \tag{15}$$

In Eq.14, it can be seen that only the convergences of ρ_t and $\|\alpha_t\boldsymbol{E}_t\|_{\mathcal{H}^\xi(\mathcal{X})}$ contribute to $\bar{r}$ while $\|u_p - f\|_{\mathcal{H}^\tau(\mathcal{X})}$ remains fixed. Compared to Eq.14, Eq.15 replaces $\|u_p - f\|_{\mathcal{H}^\tau(\mathcal{X})}$ with $\|\bar{e}_{t+1}\|_{\mathcal{H}^\tau(\mathcal{X})}$ which converges to 0. Therefore, we can conclude that $\bar{r} < \tilde{r}$. ■

4.3. Level Set Classification

Theorem 5 *Given a continuous LSE problem, assume that C2LSE with confidence parameter β reaches an ϵ-accurate solution at iteration T for any point $\boldsymbol{x} \in \mathcal{X}$ with probabilities $\bar{\delta}$ using vanilla transfer learning. Then C2LSE using AP-LSE also reaches an ϵ-accurate solution at the same iteration for any point $\boldsymbol{x} \in \mathcal{X}$ with probabilities $\tilde{\delta}$ where $\tilde{\delta} > \bar{\delta}$.*

Proof *According to Theorem 9 by Ngo et al. (2024), C2LSE with vanilla transfer learning reaches an ϵ-accurate solution for any point $\boldsymbol{x} \in \mathcal{X}$ with probability $\bar{\delta}$ if:*

$$T \geq \frac{8\beta\gamma_T}{\epsilon^2 log(1+\sigma^{-2})} \tag{16}$$

assuming $\Pr(|\bar{e}_t(\boldsymbol{x})| \leq \beta\sigma_{t-1}(\boldsymbol{x})) = \bar{\delta}, \forall \boldsymbol{x} \in \mathcal{X}, \forall t \geq 1$ *where* $\epsilon > 0$*, and* γ_T *is the maximum information gain about* f *over the set of all possible* T *noisy observations* $\mathcal{D}_T$*.*

Lemma 4 indicates that the upper bound of $\{\|\tilde{e}_t\|_{H^{\xi}(\mathcal{X})}\}_{i=1}^{\mathcal{N}^+}$ *converges to 0 faster than the upper bound of* $\{\|\bar{e}_t\|_{H^{\xi}(\mathcal{X})}\}_{i=1}^{\mathcal{N}^+}$*. This implies that for each* $\boldsymbol{x}$*, there exist* $\Delta_{\boldsymbol{x}} > 0$ *such that* $|\bar{e}_t(\boldsymbol{x})| = |\tilde{e}_t(\boldsymbol{x})| + \Delta_{\boldsymbol{x}}$*. We then have that:*

$$\begin{aligned}\Pr(|\tilde{e}_t(\boldsymbol{x})| \leq \beta\sigma_{t-1}(\boldsymbol{x})) &= \Pr(|\bar{e}_t(\boldsymbol{x})| - \Delta_{\boldsymbol{x}} \leq \beta\sigma_{t-1}(\boldsymbol{x})) \\ &< \Pr(|\bar{e}_t(\boldsymbol{x})| \leq \beta\sigma_{t-1}(\boldsymbol{x})) = \bar{\delta}, \forall \boldsymbol{x} \in \mathcal{X}, \forall t \geq 1\end{aligned} \tag{17}$$

Assuming $\Pr(|\tilde{e}_t(\boldsymbol{x})| \leq \beta\sigma_{t-1}(\boldsymbol{x})) = \tilde{\delta}, \forall \boldsymbol{x} \in \mathcal{X}, \forall t \geq 1$*, we then have* $\bar{\delta} < \tilde{\delta}$*. The remaining steps to obtain the corresponding convergence rate and level set classification accuracy are similar to both vanilla transfer learning and AP-LSE and are independent of* $\tilde{\delta}$ *and* $\bar{\delta}$*. As such, C2LSE using AP-LSE also reaches an* ϵ*-accurate solution at the same iteration for any point* $\boldsymbol{x} \in \mathcal{X}$ *with probabilities* $\tilde{\delta}$*.* ■

Theorem 5 indirectly shows a better sample efficiency for AP-LSE. It proves that by using AP-LSE, the probability that a point is classified correctly will be higher than by using vanilla transfer learning at the same iteration. Therefore, over the whole input space $\mathcal{X}$, more points will be classified correctly by using AP-LSE with the same evaluation budget. This is equivalent to AP-LSE achieving the same classification accuracy with vanilla transfer learning while using a lower evaluation budget. While the result is only proven for C2LSE, our experiments in the following section show that AP-LSE also improves the sample efficiency for other LSE algorithms.

5. Numerical Experiments

5.1. Experiment Setup

We will study the performance of AP-LSE on transfer learning for multiple LSE tasks by using it in combination with existing LSE algorithms, namely STR (the Straddle heuristic by Bryan et al. (2005)), RMILE (Zanette et al. (2019)), and C2LSE (Ngo et al. (2024)). STR is similar in spirit to the GP-UCB algorithm (Srinivas et al. (2010)), while C2LSE is a confidence-based method for continuous LSE problems. RMILE is another approach for LSE influenced by the expected improvement algorithm (Brochu et al. (2010)) and operates on discrete domains. These three algorithms comprise a comprehensive test for the effectiveness of transfer learning across different styles of LSE algorithms.

We compare AP-LSE with running LSE from scratch and vanilla transfer learning to demonstrate the sample efficiency of each method across each LSE algorithm. For a fair comparison, we start all methods without initialized data points so that the classification performance of all three methods starts similarly. Though our setting does not consider the availability of previous observations of a related LSE task (i.e. source data), we include

Diff-GP, which utilizes such data, to compare between directly using the source data like Diff-GP and using a surrogate model fitted on source data like AP-LSE. Diff-GP first models the difference between the target function's observations and the posterior mean fitted on the source data and uses that difference to adjust the source data before incorporating them into the posterior GP for the target function.

We report the average F1-score among 30 runs in each experiment. This is calculated based on the true level set labels of a fixed finite grid for each problem. Hyperparameters are selected based on the recommended values in the original works. We use a Matérn kernel with hyperparameters fitted by maximum likelihood in the section.

Synthetic Functions

We use the following synthetic functions:

- Bird: $f(x_1, x_2) = sin(x_1)e^{(1+\kappa-cos(x_2))^2} + cos(x_2)e^{(1-sin(x_1))^2} + (x_1 - x_2)^2$
- Multi-circle 3D (MC3D): $f(x_1, x_2, x_3) = e^{(sin(x_1+\kappa))^2 \times (sin(x_2+\kappa))^2 \times (sin(x_3+\kappa))^2}$
- Mishra03: $f(x_1, x_2) = \sqrt{|cos(\sqrt{x_1^2 + x_2^2} + \kappa)|} + 0.01(x_1 + x_2)$

The goal for each function is to find its superlevel or sublevel set given a function evaluation budget T. We discretize their input space into a uniform grid to evaluate the quality of level set classification. The parameter κ in these functions is to differentiate the prior and the target functions. The prior function is obtained by first setting $\kappa = 0$ and evaluating the function at T random locations. We fit a GP on these observations and use its posterior mean as the prior function for AP-LSE and vanilla transfer learning. Diff-GP directly uses these observations as its source data to simulate a similar amount of prior information between Diff-GP and the transfer learning methods (i.e. AP-LSE and vanilla transfer learning). The target function is obtained by using a different value for κ. The observation noise level is set at 0.1 for all functions. Details about each function are as follows:

	Input space	**Goal**	**Evaluation grid**	**Budget T**	**Target κ**
Bird	$\mathbf{x} \in [-6, 6]^2$	$f(\mathbf{x}) < 4$	100×100	150	0.4
MC3D	$\mathbf{x} \in [0, 6]^3$	$f(\mathbf{x}) > 1.6$	20×20×20	400	0.3
Mishra03	$\mathbf{x} \in [-5, 5]^2$	$f(\mathbf{x}) < 0.7$	100×100	250	0.4

Table 1: Details about synthetic functions

Real-world Data

We also evaluate transfer learning performance on three real-world LSE tasks. Each task includes a source dataset and a target dataset. Their details are as follows:

- Defective (Defective Area Identification (Hozumi et al. (2023))): From an active LSE perspective, the defective areas on a material surface can be estimated using only a small number of informative measurements instead of measuring all grid points which can be both slow and costly. We use the dataset provided by Hozumi et al. (2023) about red-zone identification for solar cell ingots in which the red-zone is the defective area with low carrier lifetime values. The dataset includes carrier lifetime values for two ingots (one chosen as the source dataset and the other as the target dataset),

both in the form of a 161×121 uniform grid where the value is recorded for each grid point. The target is to find the sublevel set that has carrier lifetime values under 80.

- Network (Network Latency): Inspired by Gotovos et al. (2013), we create a network latency dataset that measures the round-trip time around the world. We ping 223 servers on Dec 12th, 2023 (source dataset) and 225 servers on Dec 13th, 2023 (target dataset) to record their round-trip times. The input space includes each server's longitude and latitude coordinates. The target is to find the sublevel set that has round-trip times under 200ms indicating good internet quality.

- AA (Algorithmic Assurance): This task attempts to find scenarios in which a machine learning model's performance is at least above a certain threshold. Specifically, we train a convolutional neural network on MNIST and find its accuracy on different transformation scenarios for the test set. The input space includes continuous values of the transformations applied (i.e. rotation, scaling, and shear), and the output is the classification error on the transformed test set. Details about this task can be found in the work of Ngo et al. (2024). We construct the source dataset by training on 75% of MNIST images and the target dataset by training on all images. Each dataset includes a grid of 47,916 data points as 36×11×11×11 uniform grid. The target is to find the sublevel set that has classification error on the test set of MNIST under 4%.

Similar to the synthetic functions, a number of random data points from the source dataset, which is equal to the evaluation budget (50 for Defective, 250 for Network, and 300 for AA), are used to obtain a posterior mean serving as the prior function for AP-LSE and vanilla transfer learning and are used directly as source data for Diff-GP. The target function is obtained by selecting a random subset of data points in the target dataset (2% for Defective and AA, and 100% for Network) and then fitting a GP posterior with which the posterior mean becomes the target function.

5.2. Main Results

Figure 2 shows that AP-LSE consistently achieves better classification performances for all three functions regardless of the LSE algorithms employed. The performance gaps become smaller in the later iterations, which is expected when the transferred knowledge's effect decreases with more data observed. Similar patterns are observed in Figure 3 with better performances for AP-LSE compared to the other two transfer learning methods. For both synthetic functions and real-world problems, AP-LSE demonstrates superior classification accuracy from the beginning compared to vanilla transfer learning. This matches the goal of active LSE where high classification accuracy using few data points is desirable.

AP-LSE outperforms Diff-GP on all datasets in the long term, which indicates using an adaptive surrogate model as prior for active LSE is more effective than directly involving source data like Diff-GP even though Diff-GP takes into account the difference between the source task and the target task. We attribute this performance difference to the struggle of Diff-GP to capture the source-target difference. As noted by the authors of Diff-GP, its performance depends on the simplicity of learning this difference and the proximity of the source data to the target evaluation locations, so it struggles when the difference is too complex to model (as demonstrated in the next section). Our method does not have those

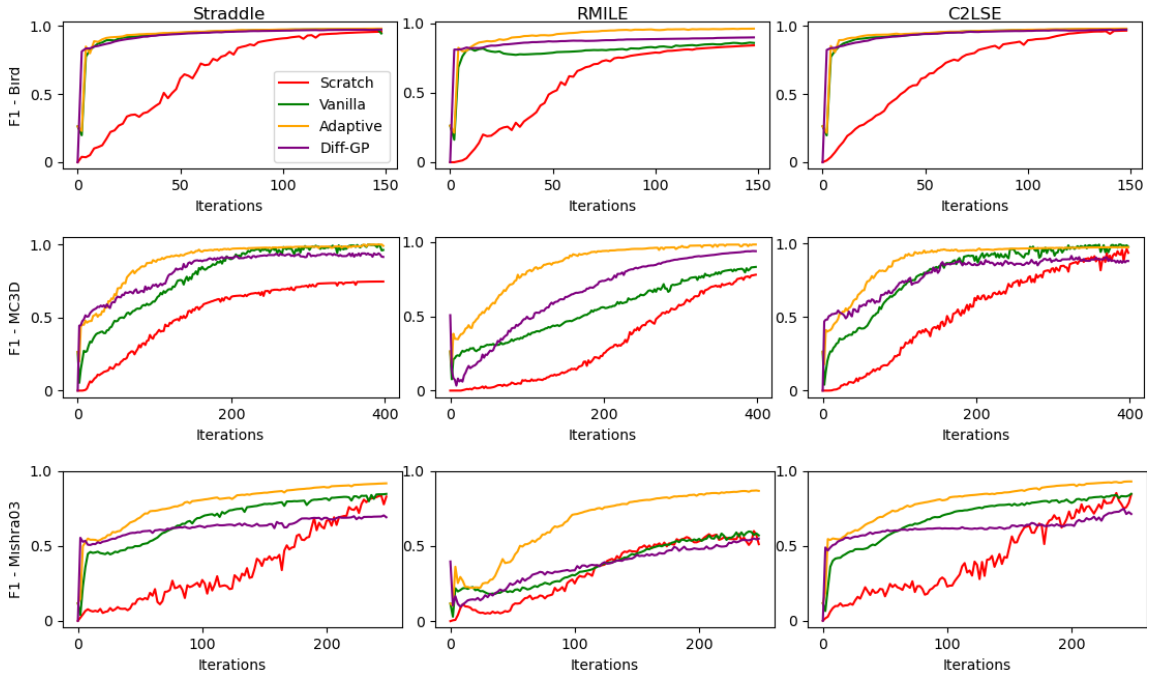

Figure 2: Transfer learning for active LSE on synthetic functions

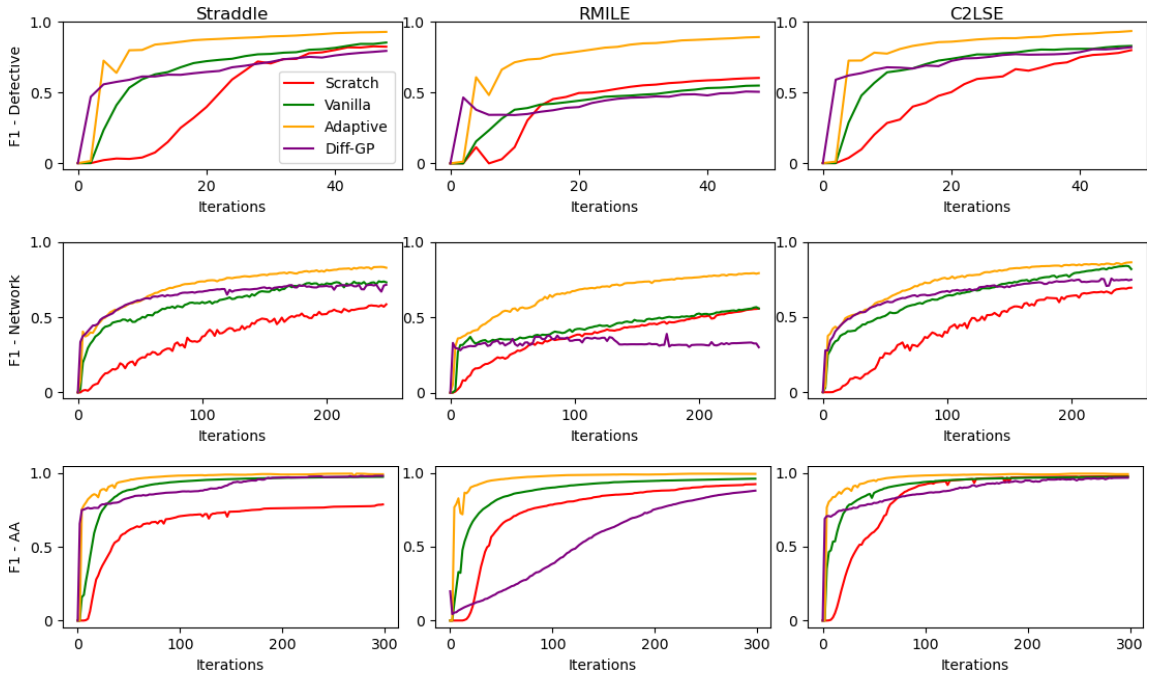

Figure 3: Transfer learning for active LSE on real-world problems

dependencies. We do not use source data and seamlessly incorporate the observed difference into our posterior rather than modeling it separately. As such, AP-LSE is not affected by the difficulty in modeling the difference between the prior and the target function.

5.3. Varying Prior Level Set Similarity

The main design goal of AP-LSE is to have robust classification performance for an LSE algorithm regardless of the prior's relevance to the target function. We now evaluates the methods on different levels of level set similarity between the prior and the target function measured by the percentage of grid points with the same level set label between the prior and the target function. We perform this experiment on Mishra03 where the prior function remains as described in Section 5.1 ($\kappa = 0$), and the target function varies by varying κ.

κ		0.2	0.4	0.6	0.8	1.0
Level set similarity		88%	77%	67%	56%	47%
Straddle	Scratch	243	229	205	**158**	**173**
	Vanilla	88	160	217	215	207
	DiffGP	59	NA	NA	NA	NA
	AP-LSE	**24**	**84**	**150**	168	**173**
C2LSE	Scratch	298	246	255	183	177
	Vanilla	78	189	250	212	198
	DiffGP	56	NA	NA	NA	NA
	AP-LSE	**20**	**85**	**154**	**170**	**165**

Table 2: #Iterations to reach 80% F1-score on Mishra03 with different similarity levels. NA means the algorithm cannot reach the target F1-score within the evaluation budget.

Table 2 shows the average number of iterations needed to reach 80% F1-score on Mishra03. On mostly all similarity levels, AP-LSE requires significantly fewer iterations than other methods, even at lower similarity. Diff-GP only works well when the difference is small as expected. It is also noteworthy that AP-LSE may not have an advantage over running from scratch if the prior function is largely different from the target function.

6. Conclusion

We proposed AP-LSE, a transfer learning method for active LSE that robustly incorporates the prior function such that an LSE algorithm can utilize the prior information while still being safeguarded from discrepancies between the prior and the target black-box function. AP-LSE locally adjusts the prior function according to the observed discrepancies to provide more accurate function estimation in level set classification. A detailed theoretical analysis shows that our method is effective in improving function estimation and consequently, in improving level set classification by increasing the classification confidence for the same accuracy. Our empirical results confirm the superior performance of AP-LSE compared to vanilla transfer learning on various active LSE tasks using different LSE algorithms.

Acknowledgments

This research was partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP210102798). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

References

Ilija Bogunovic, Jonathan Scarlett, Andreas Krause, and Volkan Cevher. Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. *NeurIPS*, 2016.

Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *ArXiv*, 2010.

Brent Bryan, Robert C. Nichol, Christopher R Genovese, Jeff Schneider, Christopher J. Miller, and Larry Wasserman. Active learning for identifying function threshold boundaries. *NeurIPS*, 2005.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. *ICML*, 2017.

Shirin Golchi. Informative priors and bayesian computation. *International Conference on Data Science and Advanced Analytics (DSAA)*, 2016.

Shivapratap Gopakumar, Sunil Gupta, Santu Rana, Vu Nguyen, and Svetha Venkatesh. Algorithmic assurance: An active approach to algorithmic testing using bayesian optimisation. *NeurIPS*, 2018.

Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. *IJCAI*, 2013.

Shota Hozumi, Kentaro Kutsukake, Kota Matsui, Syunya Kusakawa, Toru Ujihara, and Ichiro Takeuchi. Adaptive defective area identification in material surface using active transfer learning-based level set estimation. *ArXiv*, 2023.

Motonobu Kanagawa, Philipp Hennig, D. Sejdinovic, and Bharath K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *ArXiv*, 2018.

Blake Mason, Lalit Jain, Subhojyoti Mukherjee, Romain Camilleri, Kevin Jamieson, and Robert Nowak. Nearly optimal algorithms for level set estimation. *AISTAT*, 2022.

Francis Narcowich, Joseph Ward, and Holger Wendland. Sobolev error estimates and a bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 2006.

Giang Ngo, Dang Nguyen, Dat Phan-Trong, and Sunil Gupta. Active level set estimation for continuous search space with theoretical guarantee. *ACML*, 2024.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. An information-theoretic framework for unifying active learning problems. *AAAI*, 2021.

Jeremy Oakley. Eliciting Gaussian Process Priors for Complex Computer Codes. *Journal of the Royal Statistical Society Series D: The Statistician*, 51(1):81–97, 2002.

Valerio Perrone, Huibin Shen, Matthias W Seeger, Cedric Archambeau, and Rodolphe Jenatton. Learning search spaces for bayesian optimization: Another view of hyperparameter transfer learning. *NeurIPS*, 2019.

Matthias Poloczek, Jialei Wang, and Peter I. Frazier. Warm starting bayesian optimization. *Winter Simulation Conference*, 2016.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.

Shubhanshu Shekhar and Tara Javidi. Multiscale gaussian process level set estimation. *AISTAT*, 2019.

Alistair Shilton, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Regret Bounds for Transfer Learning in Bayesian Optimisation. *AISTAT*, 2017.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. *ICML*, 2010.

Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task bayesian optimization. *NeurIPS*, 2013.

Tinu Theckel Joy, Santu Rana, Sunil Gupta, and Svetha Venkatesh. A flexible transfer learning framework for bayesian optimization with convergence guarantee. *Expert Systems with Applications*, 115:656–672, 2019.

Kevin Tran, Willie Neiswanger, Kirby Broderick, Eric Xing, Jeff Schneider, and Zachary W. Ulissi. Computational catalyst discovery: Active classification through myopic multiscale sampling. *The Journal of Chemical Physics*, 154:124118, 2021.

Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme. Hyperparameter search space pruning - a new component for sequential model-based hyperparameter optimization. *ECML/PKDD*, 2015.

Andrea Zanette, Junzi Zhang, and Mykel J. Kochenderfer. Robust super-level set estimation using gaussian processes. *ECML/PKDD*, 2019.